Lecture 23

Linear Regression and Correlation

Today's Updates / Reminders

- Lab 9 is due end of the day, the last day of class.
- Homework 10 will be due end of the day, the last day of class.
- I will hold an optional exam review on Sunday, May 3, SAS1102, 2-3pm. Final exam review and key have been published. It will be Panopto'd.
- More practice resources posted to Moodle. Slides to know which statistical test is appropriate, list of symbols used, and my schedule for the end of the semester.
- ClassEval open. Closes the morning of the first day of finals. Since this is just my third semester at NC State, please help me get better by filling one out.

Example

- Moondollars is a coffee franchise that has many coffee stores all over the world. MoonDollars coffee is investigating a linear regression model to predict the annual income for a store based on the population of the city where the store is located. In the proposed regression model, annual store income is the response variable and population of the city is the explanatory variable.
- A random sample of 20 stores is selected and measurements are observed.
- Remember the trick? Right click, Copy Link and import from web in StatCrunch. Here is the data… moondollars.csv

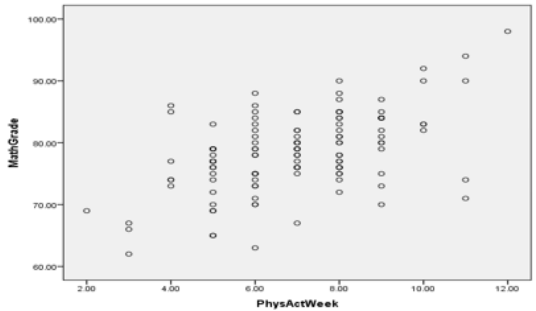
Analysis

- Let's complete all of the following:
  - Create a scatterplot
  - Find r, R^2, and the line of best fit
  - Predict the income of a store in a city of 2,120,000 and find the residual
  - Predict the income of a store in Mexico City (9.2 million). Is this valid?
  - Create a residual plot and check for linearity and constant variance
  - Do a hypothesis test at α=0.05 for correlation… $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ and StatCrunch
  - Do a hypothesis test at α=0.05 for regression.
  - Create a 95% CI for b_1, a 95% CI for mean response for 2,120, and a 95% PI for 2,120.

Is there evidence of a positive correlation between weekly time dedicated to physical activities (hours/week) and math grades for high school students? Can the math grades of high school students be predicted based on the weekly time dedicated to physical activities?
Below are the SPSS outputs of a study based on a sample of 120 high school students. For **Questions 42-44** in the next page choose the correct answer and write the corresponding letter on the line when needed.

Correlations

		PhysActWeek	MathGrade
PhysActWeek	Pearson Correlation	1	.520
	Sig. (2-tailed)		.000
	N	120	120
MathGrade	Pearson Correlation	.520	1
	Sig. (2-tailed)	.000	
	N	120	120



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.520[a]	.271	.264	5.49677

a. Predictors: (Constant), PhysActWeek

b. Dependent Variable: MathGrade

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	66.456	1.886		35.240	.000	62.722	70.190
	PhysActWeek	1.733	.262	.520	6.616	.000	1.214	2.251

a. Dependent Variable: MathGrade

42. (2 points) Provide the value of the percentage of variability in the data that is explained by the regression model. Write as a percentage with one decimal place.

Correlations

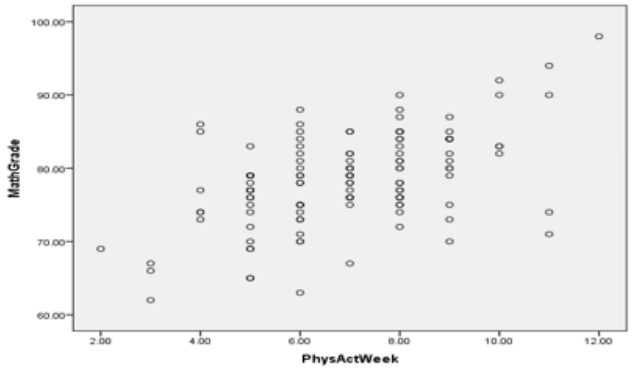
		PhysActWeek	MathGrade
PhysActWeek	Pearson Correlation	1	.520
	Sig. (2-tailed)		.000
	N	120	120
MathGrade	Pearson Correlation	.520	1
	Sig. (2-tailed)	.000	
	N	120	120

_____ 43. (2 points) Based on the output provided, we can conclude that there is evidence of a non-zero correlation between math grades and weekly time dedicated to physical activities for all high school students.

A. Because the correlation for this data set is 0.52.
B. Because the sign of the correlation for this data set is positive.
C. Because the P-value for the correlation between those two variables is 0.0.
D. Because the sample size is large enough.

_____ 44. (2 points) Using the regression equation from the SPSS outputs, if a high school student does fifteen hours of physical activities per week, her/his math grade would be 92.451.

A. This is a valid prediction as it is based on the regression equation.
B. This is an extrapolation.
C. This is an outlier.



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.520[a]	.271	.264	5.49677

a. Predictors: (Constant), PhysActWeek

b. Dependent Variable: MathGrade

4. From the coefficients output, at 5% significance level, we conclude that:

A. The population slope is positive because the lower and upper bound of the confidence interval for the slope are positive.
B. The population slope is positive because the B coefficient for PhysActWeek is 1.733.
C. The population slope is different from zero because the confidence interval for the slope doesn't contain the zero value.
D. We cannot reach any conclusion about the population slope from the coefficients output.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.520[a]	.271	.264	5.49677

a. Predictors: (Constant), PhysActWeek

b. Dependent Variable: MathGrade

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	66.456	1.886		35.240	.000	62.722	70.190
	PhysActWeek	1.733	.262	.520	6.616	.000	1.214	2.251

a. Dependent Variable: MathGrade

Multiple Regression

Chapter 13

Recap Simple Linear Regression:

- The response y depends on one explanatory variable x
- Mean of the response y → $\mu_y = \beta_0 + \beta_1 x$ (linear regression equation for the population)
- For any x, the response y is Normally distributed around this mean with a standard deviation σ that is the same for all x.
- Parameters of the model: β_0, β_1, σ

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	17.258827	5.7130261	≠ 0	65	3.0209607	0.0036
Slope	0.76100225	0.072711544	≠ 0	65	10.466044	<0.0001

$F = t^2$

Analysis of variance table for regression m

Source	DF	SS	MS	F-Stat	P-value
Model	1	12328.816	12328.816	109.53809	<0.0001
Error	65	7315.9305	112.55278		
Total	66	19644.746			

Multiple Linear Regression:

- The response y depends on p explanatory variables:

$$x_1, x_2, \ldots.. x_p$$

- Mean of the response y → $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots.. + \beta_p x_p$

(population multiple regression equation)

- For each set $x_1, x_2, \ldots.. x_p$, the response y is Normally distributed around this mean with a standard deviation σ that is the same for all the possible sets $x_1, x_2, \ldots.. x_p$.

- Parameters of the model: $\beta_0, \beta_1, \beta_2, \ldots\ldots\ldots, \beta_p, \sigma$

Data collection for multiple regression:

Simple linear regression→ n observations

$$(x_1, y_1)$$
$$(x_2, y_2)$$
$$\ldots\ldots\ldots$$
$$\ldots\ldots\ldots$$
$$(x_n, y_n)$$

Multiple linear regression:→ n observations

$$(x_{11}, x_{12}, x_{13}, \ldots\ldots\ldots, x_{1p}, y_1)$$
$$(x_{21}, x_{22}, x_{23}, \ldots\ldots\ldots, x_{2p}, y_2)$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$(x_{n1}, x_{n2}, x_{n3}, \ldots\ldots\ldots, x_{np}, y_n)$$

In software, you'll have p+1 columns, each one of length n

Estimation of the multiple regression parameters:

As in simple linear regression → Least Squares method to obtain the estimators b of the regression coefficients β.

$$b_0, b_1, ..., b_p$$

denote the estimators of the parameters

$$\beta_0, \beta_1, ..., \beta_p$$

For a given observation, the predicted response is going to be:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$$

and the residual: $e = y - \hat{y}$

→ Equation for the sample regression line

Degrees of Freedom in t-tests:

df = sample size – (# parameters needed to estimate the mean)

For 1 sample tests:

$$\mu = \mu_0 \rightarrow \text{one parameter} \rightarrow df = n\text{-}1$$

For simple linear regression:

$$\mu_y = \beta_0 + \beta_1 x \rightarrow \text{two parameters: } \beta_0, \beta_1 \rightarrow df = n\text{-}2$$

For multiple linear regression:

$$\mu_y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \rightarrow p+1 \text{ parameters: } \beta_0, \beta_1, ..., \beta_p$$

$$\rightarrow df = n - p - 1$$

Where **n** is the sample size, and **p** is the number of explanatory variables.

General procedure for Multiple Regression

- Look at the variables individually.
- Look at the variables in pairs (in particular, look at the correlations).
- Residual and Normal probability plots (to check assumptions).
- Perform a regression with all the explanatory variables.
- Interpret the results. Are all the regression coefficients significant?
- Refine the model (i.e., keep significant explanatory variables)

Software is mostly the same for both types of regression studies.
Only two differences:

- More than one explanatory variable in the independent variable box.
- Now we will lean more on the ANOVA output in the inference for the regression line!

Example: Willy Wonka's chocolate factory is getting ready to launch a new candy bar. In the process they have been advised to study the market to see how the amounts of fat, sugar and artificial colors might relate to the popularity of the candy bar. Candy bars with different amounts of fat, sugar and artificial colors are given to 30 seven-year-olds. After eating the candy bar each kid is asked to answer ten questions related to how much they liked it. A score between 0 and 60 is computed out of the ten questions.

Explanatory variables:

Response variable:

Sample size (or number of observations/cases):

Model's parameters?

Population's Regression Line?

multreg_wonka.csv

Degrees of Freedom?

Let's look at the correlations first:

To get the correlations output from StatCrunch, follow these steps:

1. Stat → Summary Stats → Correlation
2. Move variables into the box to the right
3. Tick box beside "Two-Sided P-value" then click ok.

Correlation matrix:

	Popularity	Fat	Sugar
Fat	0.5495393 (0.0017)		
Sugar	0.75575227 (<0.0001)	0.61795591 (0.0003)	
Artificial Colors	0.70423619 (<0.0001)	0.6037826 (0.0004)	0.64481228 (0.0001)

Correlation (r) between popularity and fat

P-value for HT for the correlation between popularity and fat

You can also go down to Sort rows by correlation with:

- Set column to: Popularity
- Set order to: Descending

Significance Tests for the Correlation:

Between Popularity and Fat:

$H_0 : \rho = 0$ (Popularity and Fat are independent)

$H_a : \rho \neq 0$ (Popularity and Fat are not independent)

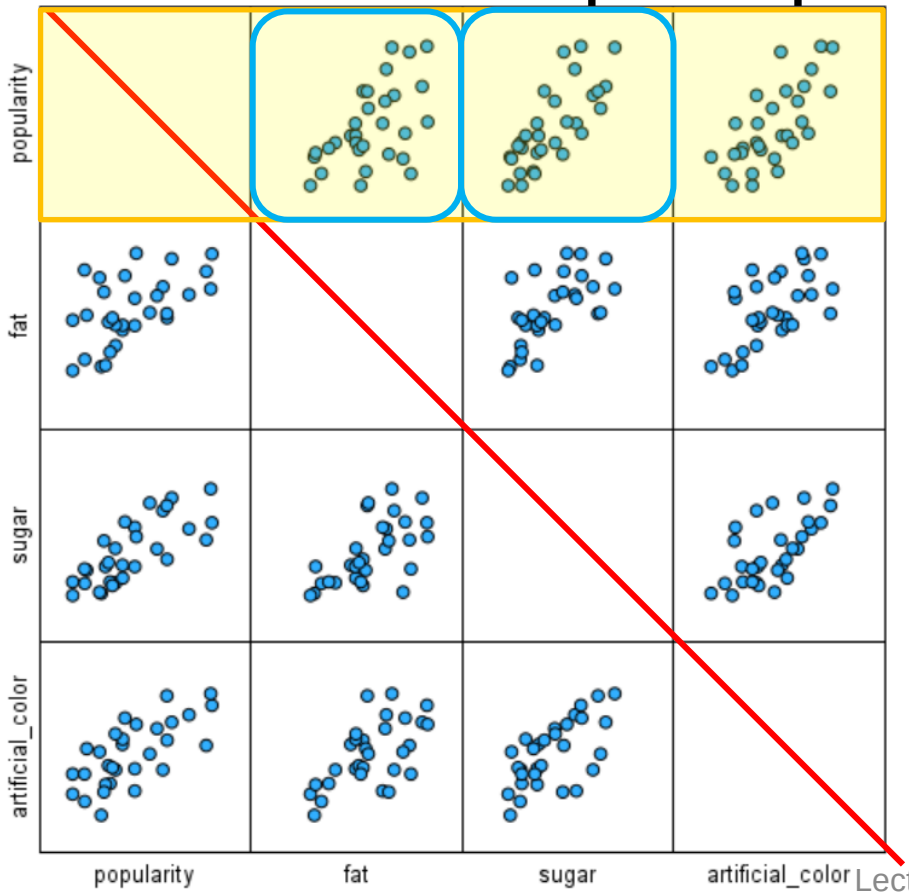
P-value=0.002<0.05=$\alpha \rightarrow$ reject the null hypothesis

We have evidence that popularity and fat **are not independent** in the population

We can test for independence of any pair of variables by looking at the corresponding P-value in the Correlations output.

We can also look at the scatterplots of any pair of variables.

Scatterplots of pairs of different variables



Between what variables do you think the correlation would be the largest?

Popularity and sugar

Between what variables do you think the correlation would be the smallest?

Popularity and fat

Graph created using SPSS:
Graphs, Scatter/Dot, Matrix Scatter

Correlations output:

SPSS:
Analyze, Correlate, Bivariate

Correlations

		popularity	fat	sugar	artificial_color
popularity	Pearson Correlation	1	.550**	.756**	.704**
	Sig. (2-tailed)		.002	<.001	<.001
	N	30	30	30	30
fat	Pearson Correlation	.550**	1	.618**	.604**
	Sig. (2-tailed)	.002		<.001	<.001
	N	30	30	30	30
sugar	Pearson Correlation	.756**	.618**	1	.645**
	Sig. (2-tailed)	<.001	<.001		<.001
	N	30	30	30	30
artificial_color	Pearson Correlation	.704**	.604**	.645**	1
	Sig. (2-tailed)	<.001	<.001	<.001	
	N	30	30	30	30

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

Correlation matrix:

	Popularity	Fat	Sugar
Sugar	0.75575227 (<0.0001)	0.61795591 (0.0003)	
Artificial Colors	0.70423619 (<0.0001)	0.6037826 (0.0004)	0.64481228 (0.0001)
Fat	0.5495393 (0.0017)		

Correlations: Strongest → Weakest

Popularity & Sugar	0.756
Popularity & Artificial colors	0.704
Sugar & Artificial colors	0.654
Artificial colors & Fat	0.604
Popularity & Fat	0.550

Correlations between response and each explanatory variable: Strongest → Weakest

Popularity & sugar	0.756
Popularity & artificial colors	0.704
Popularity & fat	0.550

Correlations between explanatory variables: Strongest and Weakest

sugar & artificial colors	0.654
artificial colors & fat	0.604

Multiple regression output

To get the outputs in StatCrunch: Stat, Regression, Multiple Linear and move all predictors to the x-variable box!

Multiple linear regression results:
Dependent Variable: Popularity
Independent Variable(s): Fat, Sugar, Artificial Colors
Popularity = -28.87677 + 0.32774129 Fat + 3.9118411 Sugar + 19.670543 Artificial Colors

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	-28.87677	19.735418	≠ 0	26	-1.4631952	0.1554
Fat	0.32774129	4.4597566	≠ 0	26	0.073488605	0.942
Sugar	3.9118411	1.2484303	≠ 0	26	3.1334077	0.0042
Artificial Colors	19.670543	8.6290548	≠ 0	26	2.279571	0.0311

R^2 = % of variation in popularity that is explained by the regression line with these explanatory variables

$$R^2 = \frac{4994.476}{7662.887}100 = 65.2\%$$

$$s = \sqrt{102.631} = 10.13$$

65.2%

Analysis of variance table for multiple regression model:

Source	DF	SS	MS	F-stat	P-value
Model	3	4994.4756	1664.8252	16.221434	<0.0001
Error	26	2668.4111	102.6312		
Total	29	7662.8867			

s = estimate for standard deviation of the model (σ)

10.1307

Summary of fit:
Root MSE: 10.130706
R-squared: 0.6518
R-squared (adjusted): 0.6116

The inference for the multiple regression line has two steps:

1. ANOVA F test for multiple regression

Aiming at writing:

$$\hat{Y}_P = b_0 + b_A x_A + b_S x_S + b_F x_F$$

where $\hat{Y}_P$ = Predicted Popularity, x_A =Artificial color, x_S =Sugar, x_F =Fat and b_0, b_A, b_S, b_F are the estimates of $\beta_0, \beta_A, \beta_S, \beta_F$

Hypothesis for ANOVA test in multiple linear regression

H_0: $\beta_A = \beta_S = \beta_F = 0$ (we are not going to look at the intercept)

H_a: Not all the βs are 0

From ANOVA output:

If P-val < α →Reject H_0→ Keep going! At least one of the coefficients is different from zero; it means there is a linear relationship in the population.

If P-val > α→ Do not reject H_0→ Stop! (It means there is no linear association in the population)

ANOVA F test for multiple regression

Model		Sum of Squares	df	Mean Square	F	Sig
1	Regression	4994.476	3	1664.825	16.221	.000(a)
	Residual	2668.411	26	102.631		
	Total	7662.887	29			

H_0: $\beta_A = \beta_S = \beta_F = 0$

H_a: Not all the βs are 0

F = 16.221, P-value = 0.000 < 0.05→ reject H_0.

We have evidence that at least one of these regression coefficients is different from zero → the response variable has a linear association with at least one explanatory variable in the population.

Source	Sum of Squares	df	Mean Square	F
Model	SSM	p	$\frac{SSM}{p}$	$\frac{MSM}{MSE}$
Error	SSE	n – (p + 1)	$\frac{SSE}{n-(p+1)}$	
Total	SST	n - 1		

$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$ $\quad R^2_{adj} = 1 - \frac{MSE}{\text{"}MST\text{"}} = 1 - \frac{MSE}{SST/N-1}$

2. Multiple regression results b→ Coefficients output

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	-28.87677	19.735418	≠ 0	26	-1.4631952	0.1554
Fat	0.32774129	4.4597566	≠ 0	26	0.073488605	0.942
Sugar	3.9118411	1.2484303	≠ 0	26	3.1334077	0.0042
Artificial Colors	19.670543	8.6290548	≠ 0	26	2.279571	0.0311

Equation for multiple regression?

What is the predicted value of popularity for 4gr of sugar, 1gr of artificial colors, and 2gr of fat?

If the value of popularity for 4gr of sugar, 1gr of artificial colors, and 2gr of fat is 6.5, find the residual:

Multiple regression results

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	-28.87677	19.735418	≠ 0	26	-1.4631952	0.1554
Fat	0.32774129	4.4597566	≠ 0	26	0.073488605	0.942
Sugar	3.9118411	1.2484303	≠ 0	26	3.1334077	0.0042
Artificial Colors	19.670543	8.6290548	≠ 0	26	2.279571	0.0311

Equation for multiple regression?

Are all the coefficients significant?→Individual t-tests

3. Individual coefficients' t-tests: t-test on each x_i takes into account that we are using x_i as well as the other variables to build the model. (df = n-p-1; n: number of data; p: number of explanatory variables) (Consider $\alpha = 0.05$)

T-test for Artificial Color

H_0: $\beta_A = 0$
H_a: $\beta_A \neq 0$
$t = 2.280$, P-value = $0.031 < \alpha$
→ Reject H_0
Evidence that Artificial Color is significant

T-test for Sugar:

H_0: $\beta_S = 0$
H_a: $\beta_S \neq 0$
$t = 3.133$, P-value = $0.004 < \alpha$
→ Reject H_0
Evidence that Sugar is significant

T-test for Fat:

H_0: $\beta_F = 0$
H_a: $\beta_F \neq 0$
$t = 0.073$, P-value = $0.942 > \alpha$
→Do not reject H_0
Evidence that Fat is NOT significant

Can we refine the model?
Can we choose one explanatory variable to leave out?

When refining the model how do we decide if eliminating an explanatory variable produced a good change or a bad change?

Good changes:

- R^2 decreases by less than 3%
- s decreases
- F increases, P-value decreases (from the ANOVA output)
- We have less number of coefficients with P-values > α

OK changes:

- R^2 decreases by less than 3%
- s increases by less than 3%

Bad changes:

- R^2 decreases by more than 3%
- s increases by more than 3%
- F decreases, P-value increases (from the ANOVA output)

When working with software we need to remove the explanatory variable that we want to eliminate (in this case, Fat) from the independent variable box (move it back to the left). Then, we rerun the multiple regression test.

StatCrunch outputs for the refined model (no Fat)

Multiple linear regression results:
Dependent Variable: Popularity
Independent Variable(s): Sugar, Artificial Colors
Popularity = -27.591815 + 3.9462673 Sugar + 19.887204 Artificial Colors

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	-27.591815	8.9818253	≠ 0	27	-3.0719608	0.0048
Sugar	3.9462673	1.1356922	≠ 0	27	3.4747685	0.0017
Artificial Colors	19.887204	7.9590091	≠ 0	27	2.4987036	0.0188

Analysis of variance table for multiple regression model:

Source	DF	SS	MS	F-stat	P-value
Model	2	4993.9213	2496.9607	25.259952	<0.0001
Error	27	2668.9654	98.850569		
Total	29	7662.8867			

Summary of fit:
Root MSE: 9.9423623
R-squared: 0.6517
R-squared (adjusted): 0.6259

Look at the changes in R^2? s? F? and P-value?
Are they bad/ok/good? Any not significant variable?

After removing the variable "Fat": Comparing the 2 models

	Original (all 3 explanatory variables)	New (only sugar and artificial color)	Change
R^2	65.2%	65.2%	None (Good)
s	10.1307	9.9424	Dropped (Good)
F, P-value	16.221, 0	25.260, 0	F increased (Good), and P-value is still very small (Good)
Insignificant explanatory variables	Fat	None	The remaining variables are significant (Good)

Regression line for refined model

Keeping just artificial color and sugar model is:

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	-27.591815	8.9818253	≠ 0	27	-3.0719608	0.0048
Sugar	3.9462673	1.1356922	≠ 0	27	3.4747685	0.0017
Artificial Colors	19.887204	7.9590091	≠ 0	27	2.4987036	0.0188

Is there any other variable that can be left out?

Using this model, predict the popularity rating for 4gr of sugar and 1gr of artificial colors

If the popularity score for 4gr of sugar and 1gr of artificial color is 6.5, find the residual:

Checking for assumptions: Residuals and Normality Plots

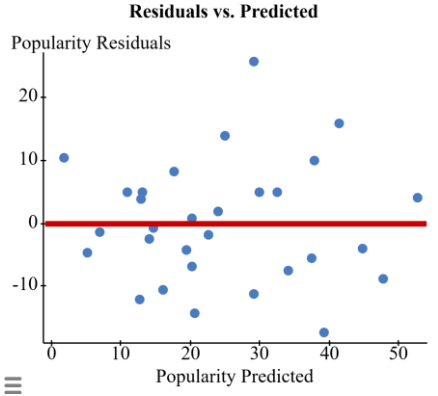
What kind of plot will let us see if the model assumptions are met?

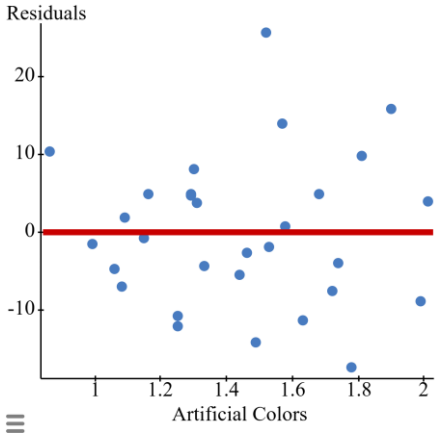
- **Scatterplot (y vs. x)**: Let us **check** if $\mu_y=\beta_0+\beta_1 x$. If we see the data are aligned in the scatterplot (**linear pattern**) → assumption 3 is established.

- **Residual plot ((y-ŷ) vs. x):** Let us **check linearity** (values spread around zero) and if different **values of y for a given x are independent** (assumption 1). Let us see if the we could have the **same σ** (for example, if the spread of data around 0 line is similar for all values of x, that indicates assumption 4 is met)

- **Normal probability plot for the residuals:** **check** for assumption 2: y **Normally distributed**. The residuals are also Normally distributed with mean equal zero. If the Normal probability plot for the residuals shows a linear behavior → Normality.

Residual plots and Normal probability plot for the refined model

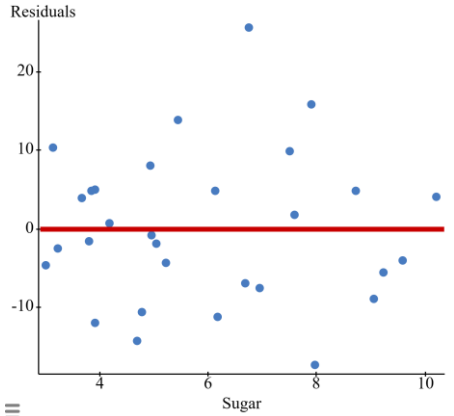


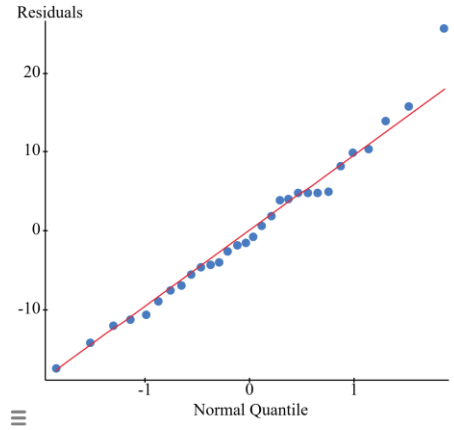


In StatCrunch, top left plot comes from Graphs dialog box, Residuals vs. Predicted.

For others, generate the residuals. Under Save dialog box, select Residuals.

Then do Graph, Scatterplot and Graph, QQ Plot





Example

- What is a good predictor for a final exam score?
  - Homework average?
  - Previous exams?
  - PRWA?
  - Labs?
- We can run multiple regression and see what we find!

multreg_grades_370spfa25.csv

72. An entomologist is interested in the longevity of fruit flies. He thinks the length of the thorax and the percentage of each day spent sleeping might affect longevity. Here are the variables:

Longevity Lifespan, in days

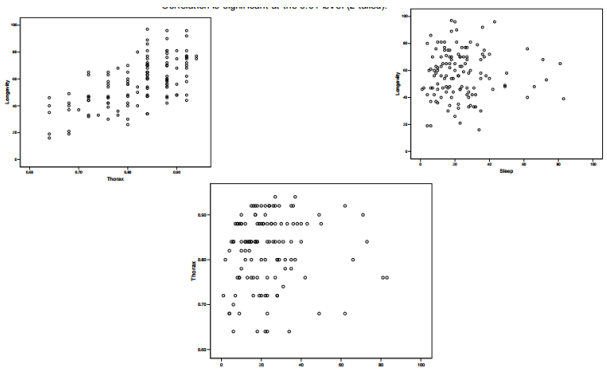
Thorax Length of thorax, in mm (x.xx)

Sleep Percentage of each day spent sleeping

Correlations

		Longevity	Thorax	Sleep
Longevity	Pearson Correlation	1	.636**	.004
	Sig. (2-tailed)	.	.000	.963
	N	125	125	125
Thorax	Pearson Correlation	.636**	1	.066
	Sig. (2-tailed)	.000	.	.466
	N	125	125	125
Sleep	Pearson Correlation	.004	.066	1
	Sig. (2-tailed)	.963	.466	.
	N	125	125	125

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



SPSS output for using THORAX and SLEEP to predict LONGEVITY:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.638[a]	.407	.397	13.641

a. Predictors: (Constant), Sleep, Thorax

b. Dependent Variable: Longevity

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15551.370	2	7775.685	41.787	.000[a]
	Residual	22701.430	122	186.077		
	Total	38252.800	124			

a. Predictors: (Constant), Sleep, Thorax

b. Dependent Variable: Longevity

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-60.533	13.076		-4.629	.000	-86.419	-34.646
	Thorax	144.899	15.850	.639	9.142	.000	113.522	176.277
	Sleep	-.042	.077	-.038	-.542	.589	-.195	.111

a. Dependent Variable: Longevity

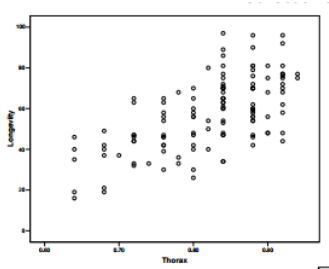
72. An entomologist is interested in the longevity of fruit flies. He thinks the length of the thorax and the percentage of each day spent sleeping might affect longevity. Here are the variables:

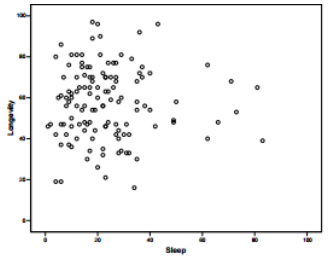
Longevity	Lifespan, in days
Thorax	Length of thorax, in mm (x.xx)
Sleep	Percentage of each day spent sleeping

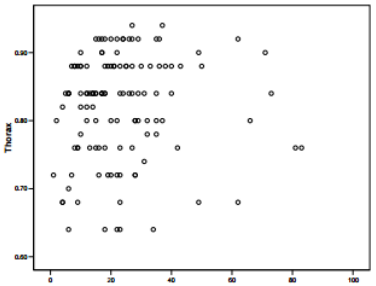
1. Name the explanatory variables and the response variable.

   Explanatory variables: Response:

2. According to the scatterplots, the pair of variables that shows the strongest correlation is:
   A. Longevity & Sleep
   B. Longevity & Thorax
   C. Thorax & Sleep







To see if two variables are independent in the population we need to look at:

A. Correlations output
B. Model Summary
C. ANOVA output
D. Coefficients output

Correlations

		Longevity	Thorax	Sleep
Longevity	Pearson Correlation	1	.636**	.004
	Sig. (2-tailed)	.	.000	.963
	N	125	125	125
Thorax	Pearson Correlation	.636**	1	.066
	Sig. (2-tailed)	.000	.	.466
	N	125	125	125
Sleep	Pearson Correlation	.004	.066	1
	Sig. (2-tailed)	.963	.466	.
	N	125	125	125

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

SPSS output for using THORAX and SLEEP to predict LONGEVITY:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.638[a]	.407	.397	13.641

a. Predictors: (Constant), Sleep, Thorax

b. Dependent Variable: Longevity

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15551.370	2	7775.685	41.787	.000[a]
	Residual	22701.430	122	186.077		
	Total	38252.800	124			

a. Predictors: (Constant), Sleep, Thorax

b. Dependent Variable: Longevity

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-60.533	13.076		-4.629	.000	-86.419	-34.646
	Thorax	144.899	15.850	.639	9.142	.000	113.522	176.277
	Sleep	-.042	.077	-.038	-.542	.589	-.195	.111

a. Dependent Variable: Longevity

Regarding the population we can conclude:

A. There is evidence that Longevity & Thorax are not independent in the population because Sig.=0
B. There is no evidence that Longevity & Thorax are not independent in the population because Sig.=0
C. There is evidence that Longevity & Thorax are not correlated in the population because Sig.=0
D. There is no evidence that Longevity & Thorax are correlated in the population because Sig.=0

There is evidence that the correlation between Longevity & Thorax is:

A. 0.636 in the sample
B. 0.636 in the population

Correlations

		Longevity	Thorax	Sleep
Longevity	Pearson Correlation	1	.636**	.004
	Sig. (2-tailed)	.	.000	.963
	N	125	125	125
Thorax	Pearson Correlation	.636**	1	.066
	Sig. (2-tailed)	.000	.	.466
	N	125	125	125
Sleep	Pearson Correlation	.004	.066	1
	Sig. (2-tailed)	.963	.466	.
	N	125	125	125

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

SPSS output for using THORAX and SLEEP to predict LONGEVITY:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.638[a]	.407	.397	13.641

a. Predictors: (Constant), Sleep, Thorax

b. Dependent Variable: Longevity

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15551.370	2	7775.685	41.787	.000[a]
	Residual	22701.430	122	186.077		
	Total	38252.800	124			

a. Predictors: (Constant), Sleep, Thorax

b. Dependent Variable: Longevity

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-60.533	13.076		-4.629	.000	-86.419	-34.646
	Thorax	144.899	15.850	.639	9.142	.000	113.522	176.277
	Sleep	-.042	.077	-.038	-.542	.589	-.195	.111

a. Dependent Variable: Longevity

6. The small significance (P-value) in the ANOVA output indicates:

   A. There is evidence that all the coefficients in the population's regression model are different from zero.
   B. There is evidence that at least one coefficient in the population's regression model is different from zero.

7. To write down the equation for the regression line we need to look at:

   A. Correlations output
   B. Model Summary
   C. ANOVA output
   D. Coefficients output

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-60.533	13.076		-4.629	.000	-86.419	-34.646
	Thorax	144.899	15.850	.639	9.142	.000	113.522	176.277
	Sleep	-.042	.077	-.038	-.542	.589	-.195	.111

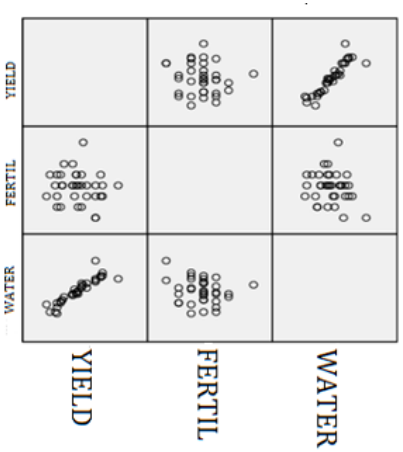
a. Dependent Variable: Longevity

8. Write down the regression model for the predicted Longevity in terms of Thorax and Sleep:

9. Choose the most appropriate answer; assume 5% significance level.
   A. There is no need to drop any explanatory variable since both are significant.
   B. I would choose to rerun the model without Thorax.
   C. I would choose to rerun the model without Sleep.

Can we predict the pounds of peaches produced by a tree ("YIELD") based on the amount of fertilizer (in lbs) per application ("FERTIL") and the gallons of water it receives per week during the summer months ("WATER")?

Below are the SPSS outputs for the analysis of this problem. For the questions 17-22 in the next page choose the correct answer and write the corresponding letter on the line when needed.



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.914[a]	.836	.825	4.162

a. Predictors: (Constant), WATER, FERTIL
b. Dependent Variable: YIELD

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2815.413	2	1407.706	81.284	.000[b]
	Residual	554.187	32	17.318		
	Total	3369.600	34			

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	99.422	5.786		17.182	.000	87.636	111.209
FERTIL	.304	.504	.044	.603	.551	-.722	1.330
WATER	1.132	.090	.922	12.610	.000	.949	1.315

16. (2 points) Write down the names of the response variable and explanatory variable(s):

_____17. (2 points) In between what variables is the correlation the strongest?

A. YIELD and WATER.
B. YIELD and FERTIL.
C. WATER and FERTIL.

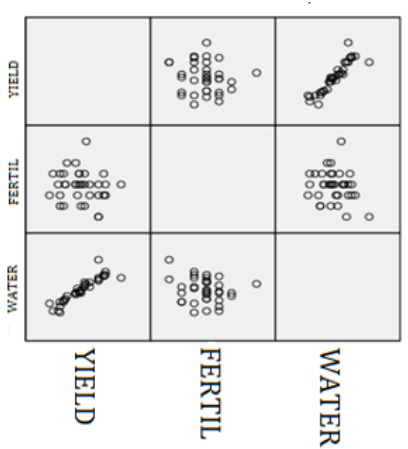
_____18. (2 points) The percentage of variation in the response explained by the regression model is:

A. 91.4%
B. 83.6%
C. 82.5%
D. 81.284%

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.914[a]	.836	.825	4.162

a. Predictors: (Constant), WATER, FERTIL
b. Dependent Variable: YIELD



19. (4 points) Write down the regression equation for predicting the production of peaches per tree based on the amount of fertilizer per application and the gallons of water it receives per week during the summer months. Make sure you explicitly state the different variables involved in the equation (3 points for correct coefficients, 1 point for correct notation).

20. (2 points) A peach tree receives 5 lbs of fertilizer per application and 39 gallons of water per week during the Summer months. Find the predicted production (in lbs) using the full regression equation from the previous question, **round to 2 decimal places**.

21. (2 points) The data collected show the same peach tree (as in previous question) produced 148 lbs of peaches. Calculate the residual for this peach tree, **round to 2 decimal places**.

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	99.422	5.786		17.182	.000	87.636	111.209
FERTIL	.304	.504	.044	.603	.551	-.722	1.330
WATER	1.132	.090	.922	12.610	.000	.949	1.315

_____22. (3 points) Choose the correct answer:

A. The regression model including all the explanatory variables is the best because Sig. for ANOVA is zero.

B. The variables "Constant" and "WATER" might be removed to improve the model because their Sig. is zero.

C. The variable "FERTIL" might be removed to improve the model because Sig. is 0.551.

ANOVA[a]

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	2815.413	2	1407.706	81.284	.000[b]
Residual	554.187	32	17.318		
Total	3369.600	34			

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	99.422	5.786		17.182	.000	87.636	111.209
FERTIL	.304	.504	.044	.603	.551	-.722	1.330
WATER	1.132	.090	.922	12.610	.000	.949	1.315