Baseline Professional Learning Handout

M214: Statistical Error and Predictive Modeling Validation

Back to Table of Contents

Selected Common Core High School Standards for Mathematics

Summarize, represent, and interpret data on two categorical and quantitative variables

5. Summarize categorical data for two categories in two-way frequency tables. Interpret relative frequencies in the context of the data (including joint, marginal, and conditional relative frequencies). Recognize possible associations and trends in the data.

6. Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.
   a. Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear, quadratic, and exponential models.
   b. Informally assess the fit of a function by plotting and analyzing residuals.
   c. Fit a linear function for a scatter plot that suggests a linear association.

Back to Table of Contents

M214 Content and Practice Expectations

214.a	Demonstrate understanding of the role of variability in predictive modeling.
214.b	Use Pearson's r to evaluate predictive models.
214.c	Use residuals to assess model fit and decide if changes to the predictive model or its variables are necessary.
214.d	Compare predictive models using statistical techniques in order to improve these models and decide which model makes the best predictions.
214.e	Consider the bias-variance trade-off that occurs when making a predictive model.

Back to Table of Contents

M214 Content and Practice Expectations and Indicators

Content and Practice Expectations	Indicators Choose an artifact where you...
214.a: Demonstrate understanding of the role of variability in predictive modeling.	i. use technology to calculate measures of variability commonly used in the field of Data Science, e.g., Sum of Squares, Mean Absolute Error, variance, standard deviation, etc. and interpret the meaning of these measures in the context of the data set(s) being used.
	ii. demonstrate understanding that measures of variability can be used to assess predictive model fit.
	iii. demonstrate understanding that predictive modeling is a way to explain variation of a response.
214.b: Use Pearson's r to evaluate predictive models.	i. interpret Pearson's r in context and understand its purpose in quantifying the strength of a linear relationship.
	ii. demonstrate understanding that a correlation coefficient close to -1 or 1 does not necessarily mean that a linear model is appropriate.
	iii. anticipate the value of Pearson's r between two variables by visually assessing scatterplots before calculating its value using technology.
	iv. demonstrate understanding of the effect outliers have on the value of Pearson's r.
	v. calculate the value of R^2 using technology and interpret in context.
214.c: Use residuals to assess model fit and decide if changes to the predictive model or its variables are necessary.	i. use technology to produce residual summary statistics and residual plots for data sets and use them to assess model fit.
	ii. recognize outliers, patterns, random scatter, or heteroskedasticity in residual plots or the standard deviation of residuals to make decisions about predictive model fit, e.g., transforming variables to improve fit.
214.d: Compare predictive models using statistical techniques in order to improve these models and decide which model makes the best predictions.	i. use one or more statistical techniques to check for improved predictive model fit, e.g., F-tests, sensitivity analysis, transforming variables, adding or excluding variables, ANOVA, residual plots, normalizing variables, cross validation, etc.
	ii. compare measures of variability and Pearson's r of two or more predictive models to decide which model makes the best prediction.
	iii. use an iterative process to improve a predictive model.
	iv. make predictions using the improved predictive model and interpret these predictions in context.

Back to Table of Contents

Content and Practice Expectations	Indicators Choose an artifact where you...
214.e: Consider the bias-variance trade-off that occurs when making a predictive model.	i. demonstrate understanding of bias-variance trade-off, namely that a model with high bias tends to have low variance and a model with high variance tends to have low bias.
	ii. make decisions about the complexity of a predictive model after weighing the bias-variance trade-off within the model.

Let's Investigate

Back to Table of Contents

Used Subaru Foresters I

Jane wants to sell her Subaru Forester, but doesn't know what the listing price should be. She checks on craigslist.com and finds 22 Subarus listed. The table below shows age (in years), mileage (in miles), and listed price (in dollars) for these 22 Subarus. (Collected on June 6th, 2012 for the San Francisco Bay Area.)

Age	Mileage	Price
8	109428	12995
5	84804	14588
3	55321	20994
3	57474	18991
1	11696	19981
13	125260	6888
10	67740	9888
11	97500	6950
6	36967	19700
12	148000	3995
2	29836	18990
3	32349	21995
10	161460	5995
4	68075	12999
3	30007	22900
8	66000	13995
10	93450	8488
3	35518	22995
3	30047	20850
8	107506	11988
11	89207	8995
13	141235	5977

a. Make appropriate plots with well-labeled axes that would allow you to see if there is a relationship between price and age and between price and mileage. Describe the direction, strength and form of the relationships that you observe. Does either mileage or age seem to be a good predictor of price?
b. If appropriate, describe the strength of each relationship using the correlation coefficient. Do the values of the correlation coefficients agree with what you see in the plots?
c. Pick the stronger relationship and use the data to find an equation that describes this relationship. Make a residual plot and determine if the model you chose is a good one. Write a few sentences explaining why (or why not) the model you chose is appropriate.
d. If Jane's car is 9 years old with 95000 miles on it, what listing price would you suggest? Explain how you arrived at this price.

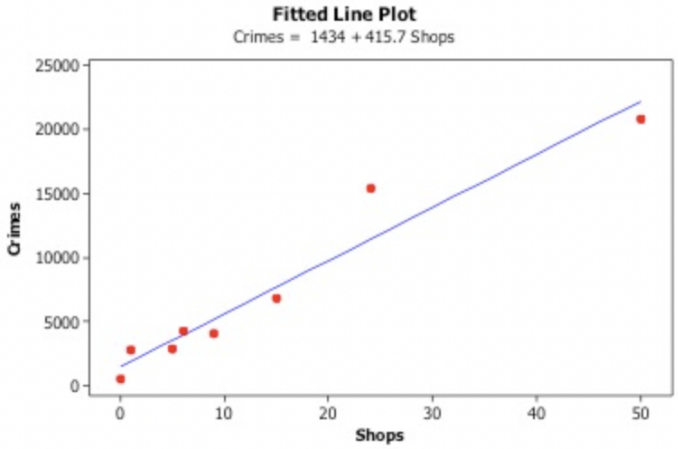
Source: Illustrative Mathematics

Back to Table of Contents

Coffee and Crime

Many counties in the United States are governed by a county council. At public county council meetings, county residents are usually allowed to bring up issues of concern. At a recent public County Council meeting, one resident expressed concern that 3 new coffee shops from a popular coffee shop chain were planning to open in the county, and the resident believed that this would create an increase in property crimes in the county. (Property crimes include burglary, larceny-theft, motor vehicle theft, and arson -- From http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/property-crime accessed on December 5, 2012.)
To support this claim, the resident presented the following data and scatterplot (with the least-squares line shown) for 8 counties in the state:

County	Shops	Crimes
A	9	4000
B	1	2700
C	0	500
D	6	4200
E	15	6800
F	50	20800
G	5	2800
H	24	15400



The scatterplot shows a positive linear relationship between "Shops" (the number of coffee shops of this coffee shop chain in the county) and "Crimes" (the number of annual property crimes for the county). In other words, counties with more of these coffee shops tend to have more property crimes annually.

a. Does the relationship between Shops and Crimes appear to be linear? Would you consider the relationship between Shops and Crimes to be strong, moderate, or weak?
b. Compute the correlation coefficient. Does the value of the correlation coefficient support your choice in part (a)? Explain.

Back to Table of Contents

c. The equation of the least-squares line for these data is: Predicted Crimes=1434+415.7(Shops) Based on this line, what is the estimated number of additional annual property crimes for a given county that has 3 more coffee shops than another county?
d. Do these data support the claim that building 3 additional coffee shops will necessarily *cause* an increase in property crimes? What other variables might explain the positive relationship between the number of coffee shops for this coffee shop chain and the number of annual property crimes for these counties?
e. If the following two counties were added to the data set, would you still consider using a line to model the relationship? If not, what other types (forms) of model would you consider?

County	Shops	Crimes
I	25	36900
J	27	24100

Source: Illustrative Mathematics

Back to Table of Contents

Let's do some math! Round 1

Task 1: Batting Average and Wins

Build a bivariate model using machine learning relating variables in a baseball data set. While running Train/Test/Split in Colab, you will explore how model complexity affects the predictability of the algorithm. Run the program using Batting Average as the input target variable and Wins as the output predictor variable to answer the following questions:

- Run the cell with a complexity of 1. What function does this create?
- What is the difference between train data and test data?
- What do the points represent? What are the inputs and outputs of this model?
- Which model do you feel would be most predictable?
- Does lower testing error or lower training error indicate a more predictable model?
- How would you use this model to make a prediction?
- What does the ERR value mean?
- How would you describe the complexity of this model?
- Does increase in complexity improve the model? What complexity value do you feel is the best choice?

Sample Learning Experience

Students can choose different variables to plot against each other and vary the complexity level to see how it affects trends and errors in train and test data.

Classroom discussion points:

- Which variables are you interested in? When you create bivariate models, are there any surprises in the given context? (ex: players with fewer home runs may have a greater salary—are there other factors that impact salary?)
- What does complexity value represent? When you increase the complexity, how does this affect the trend line? Does this increase predictability?

Adapted from YouCubed

Note: Lessons from the YouCubed.org website may require setting up a free account and logging in to access the whole lesson.

Back to Table of Contents

Learning Principles

Engage with cognitively demanding tasks in heterogeneous settings (LP 1). Students should be given opportunities to grapple with multistep, non-routine tasks that promote mathematical rigor. These experiences should be differentiated so that all students engage in appropriate challenges—for example, through tasks with multiple entry points and solution pathways. These experiences should continue to integrate knowledge and skills developed in grades 6-8 at the level of sophistication of high school mathematics.

Engage in social activities (LP 2). Students should have opportunities to work independently and to communicate with one another about mathematics by engaging in collective and collaborative learning activities. Explaining and having opportunities to revise one's thinking has excellent value in solidifying one's knowledge.

Build conceptual understanding through reasoning (LP 3). Students should be given the opportunity to reason, justify, and problem solve with critical thinking, reading, writing, speaking, and listening. Through reasoning and work with multiple representations, students learn why procedures work and build a conceptual understanding of key mathematical ideas.

Have agency in their learning (LP 4). Students should be able to choose tasks and learning experiences that align with their interests and aspirations. All students have rich and varied experiences and home lives. Learning mathematics should bring students' identities and interests to the fore and build on the strengths that they bring to the learning space.

View mathematics as a human endeavor across centuries (LP 5). Students should understand that mathematical ideas emanated over time from civilizations around the world and have opportunities to explore these contributions to mathematics. Students should develop an appreciation of mathematics as a human endeavor: one in which they feel a sense of belonging, where they see themselves as mathematicians, and one that offers opportunities to broaden their ideas about what mathematics is, how it is used, and who it is for.

See mathematics as relevant (LP 6). Students should engage with mathematics in ways that authentically involve real-world situations. Problem-solving contexts should allow students to perceive mathematics as a tool for addressing the questions that arise in everyday life, as well as the ways it can model our world and address global economic, social, and environmental challenges. Students should also engage with mathematics in ways that connect to academic disciplines and future careers by doing the mathematics used by artists, designers, engineers, and other professionals.

Employ technology as a tool for problem-solving and understanding (LP 7). Research indicates that technology is a powerful tool for learning deeper mathematics by improving calculation efficiency and enabling more sophisticated analysis. Students should learn to use technology, with emphasis put on widely used tools and software, such as calculators and spreadsheets, to make sense of models. Technology use should not be limited to supporting "doing mathematics," but should also be used as a tool for displaying and communicating results to appropriate audiences.

Back to Table of Contents

Task 1: Batting Average and Wins

Example 1:

Student work goes here

Example 2:

Student work goes here

Back to Table of Contents

Let's do some math! Round 2

Task 2: iPhone Sales

Here is data of iPhone sales during the opening weekends:

iPhone	Year	Units Sold (millions)
Original	2007	0.5
3G	2008	1
3Gs	2009	1
4	2010	1.7
4S	2011	4
5	2012	5
5C, 5S	2013	9
6, 6 Plus	2014	10
6S, 6S Plus	2015	13

Work with a partner to complete the following questions. Be prepared to share your findings with another group:

- Use technology to create a scatterplot of the data with year as the explanatory variable and units sold as the response.
- Describe the form of the relationship.
- Use the technology to find the least squares regression line, and include its graph on your scatterplot.
- Use the least squares regression line to calculate the residual for 2007. Interpret the residual.
- Now graph the residuals using technology.
- For which points was the actual greater than the predicted? For which was it less than predicted? Identify these on your graph.
- Is the regression line a good fit for the data? Why or why not? Explain using the residual plot.
- Use various transformations of Year and/or Units Sold to check if there is a better model that 2 explains the relationship between Year and Units Sold.

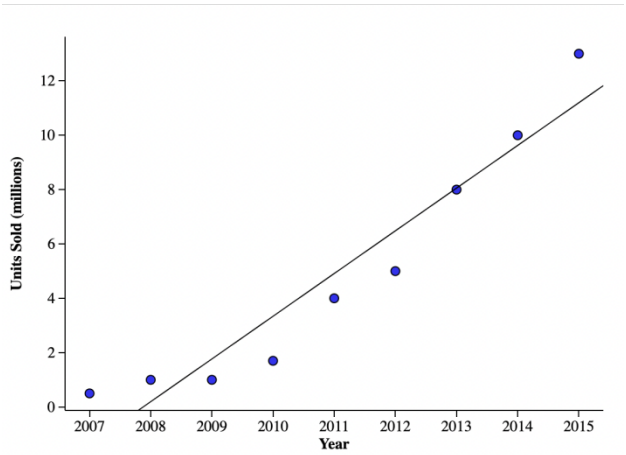
Sample Learning Experience

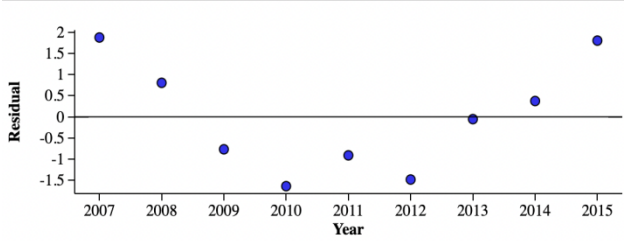
Before beginning the activity, pose the following question to students: "How does the computer or calculator find the line of 'best' fit? What makes it the line of best fit?" Using this Desmos eTool, let students try to move the line into the "best" spot (be sure you hide the line of best fit to start). Help them discover that the line of "best" fit is the one that minimizes the sum of the squared residuals (the least squares regression line). This can also be done nicely using the statistical software Fathom. Now students are ready for the activity!

When asked if a linear model is appropriate, students will sometimes use only the correlation value, r, to justify linearity. However, a strong correlation value doesn't mean an association is linear. An association can be clearly nonlinear and still have a correlation close to ±1. Only a residual plot can adequately address whether a line is an appropriate model for the data by showing the pattern of deviations from the line. For example, graphing the function $y = x^2$ for the integers 1 to 10 yields a correlation of r = 0.97, but the residual plot shows an obvious pattern.

Back to Table of Contents

Students will produce the following graphs in pairs:





Have students answer the questions, then share their findings with another pair. After some time, discuss findings as a class.

Classroom discussion points:

- What do you notice about your two graphs? What do you wonder?
- Is there a type of model that would fit this data better than a line?
- What modifications can we make to our LSRL model to better fit this data?
- What similarities and differences do these graphs have?
- What does the residual plot tell us about the relationship between Year and Units Sold?
- What types of two-variable relationships appear curved?
- How will we know when a better relationship is accomplished?
- What is the value of Pearson's r in your original LSRL? Does this value fully explain the strength of the linear relationship? Why or why not?

Adapted from StatsMedic

Back to Table of Contents

Task 2: iPhone Sales

Example 1:

Student work goes here

Example 2:

Student work goes here

Back to Table of Contents

Criteria for Success

Conference and Provide Revision Support	**Accept with Revision**	**Accept**
The student's artifact shows an emerging understanding of the expectations of the indicator(s). After conferencing and additional instruction/learning, the student may provide a revised or different artifact as evidence of the indicator(s).	The student's artifact is approaching a full understanding of the expectations of the indicator(s). The artifact may contain execution errors that should be corrected in revision. The student may revise the selected artifact or submit a different artifact.	The student's artifact demonstrates evidence that they have met the expectations of the indicator(s).

Back to Table of Contents