Linear and logistic regressions: What they try to explain, and how to interpret the results

William F. Sensakovic, PhD, DABR, MRSC
Co-Founder, Telerad Physics Teaching, LLC
Medical Physicist – Florida Hospital
Assoc. Prof. – University of Central Florida
Clin. Asst. Prof. – Florida State University
Adj. Prof. – Adventist University

Twitter: @wfsensak
E-mail: wfsensak@gmail.com
www.teleradphysics.com

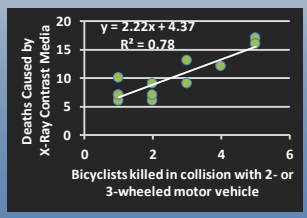
FLORIDA HOSPITAL
The skill to heal. The spirit to care.

Relevant Conflicts of Interest

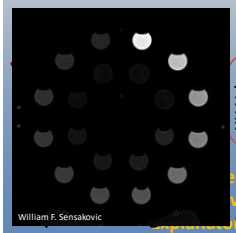
- Nope, I make no money off of teaching or performing stats

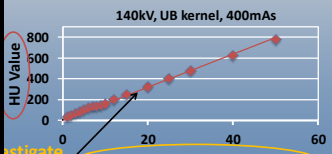
Simple Linear Regression

- After this you should be able to:
  - Understand basic terminology
  - Interpret results of a regression analysis
  - Determine if a model is "good"
  - Avoid common pitfalls



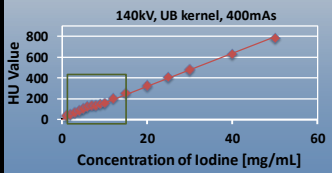
Terminology





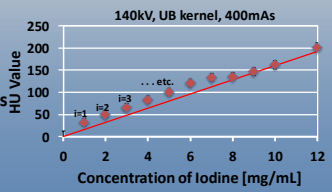
estigate
variable
determine the best fit line to the data
- Independent variable

Model Terminology



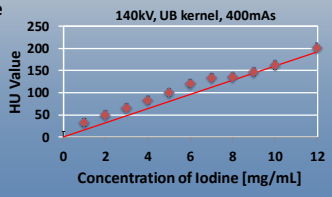
Model Terminology

- Model = Line
- Y=mX + b . . . or
- $Y_i=\beta_1 X_i + \beta_0 + \epsilon_i$
  - Y_i: Outcome value i
  - β_1: Slope that defines outcome-predictor relationship
  - X_i: Predictor value i
  - β_0: Intercept
  - ϵ_i: Point i residual error



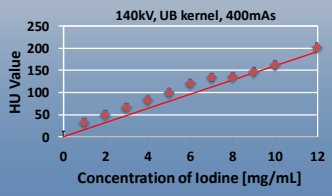
Model Terminology

- The measured HU value at any point is the best fit line plus the residual error
- Our best guess of HU at unmeasured concentrations are the values of the best fit line



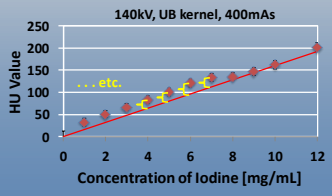
What does the model mean?

- $Y_i = \beta_1 X_i + \beta_0 + \epsilon_i$
  - Y = 16X
- β_1 (slope)
  - Every 1mg/mL of I increases HU by 16
- β_0 (intercept)
  - 0HU (water) at 0mg/mL of I
- ϵ is the residual error
  - Notice smaller above 8mg/mL



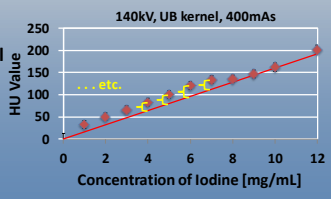
What does "best" mean?

- Point Residual Error
  - Point-to-Line Distance
- Residual Error
  - Combine point residual errors together



What does "best" mean?

- "Best" minimizes the sum of squared residual errors (SSE)
  - Squared since distance can be positive or negative

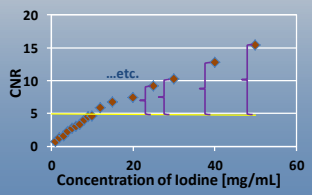


How good is our model?

- Not a simple question – several things to look at
- Coefficient of Determination (r^2 or R^2)
  - Essentially answers how much of the variability can be removed by the data
  - Best explained through example
    - CNR vs Concentration of Iodine

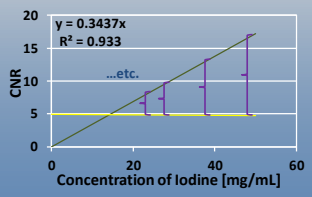
Coefficient of Determination

- Total Sum of Squares
  - How much does the data differ from the mean?
  - Variation in the data from all sources



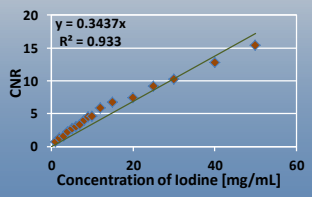
Coefficient of Determination

- Regression Sum of Squares
  - How much do model-predicted values differ from the mean?
  - Variation of the model



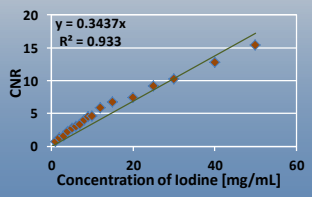
Coefficient of Determination

- How much variation does the model explain?
  - $R^2 = \frac{Model\ Variation}{Total\ Data\ Variation}$
    - 0: accounts for no variation
    - 1: perfect fit



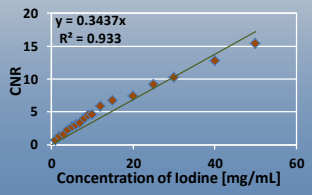
Coefficient of Determination

- How much variation does the model explain?
  - R^2 = 0.933 or ~93%
    - So not bad, but does that mean it is the correct model . . . Maybe



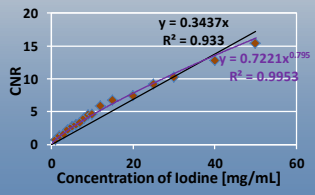
Coefficient of Determination

- What about the other 7%?
  - Random?
    - Measurement error
  - Other predictor?
  - Wrong model?



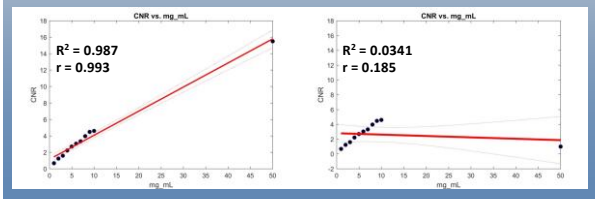
Coefficient of Determination

- Linear model
  - $R^2 = 0.933$
- Power model
  - $R^2 = 0.9953$
- Is power model the correct fit?
  - . . . maybe
- R^2 just tells you how much variation the model accounts for, nothing more

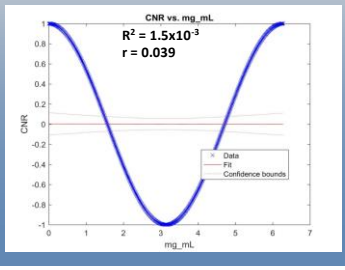


Disproportionate Impact of Outliers/Leverage Points

– Especially if far from other data



No Linear Relationship ≠ No Relationship



Hypothesis Testing

- Is there a relationship between CNR and image contrast?
  - Is slope significant?
    - Null (H_0): $\beta_1=0$
    - Alternate (H_1): $\beta_1 \neq 0$

	Estimate	SE	tStat	pValue
mg_mL	0.34366	0.013239	25.958	1.662e-14

Hypothesis Testing

- Estimate of $\beta_1 = 0.34$
- Standard Error of β_1 Estimate = 0.013
- Estimate is ***26x*** the SE so we get a very small p-value and "confirm" the linear model
  - Does NOT mean a curvilinear model is not better

	Estimate	SE	tStat	pValue
mg_mL	0.34366	0.013239	25.958	1.662e-14

Pitfall

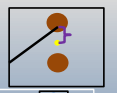
- Statistical significance does not imply clinical significance
- A model may have a significant slope, but explain <1% of variation
  - Probably not a useful model

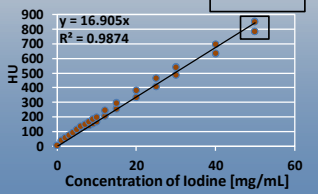
Lack-of-Fit Test

- If residual error due to the model is large compared to residual error from other sources, then we have a poor model
  - Requires repeated outcome values for the same predictor value (i.e., replicates)

Lack-of-Fit Test

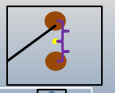
- 2 replicates
- Lack of fit sum of squares
  - Squared sum of distances between mean of data values and model at i

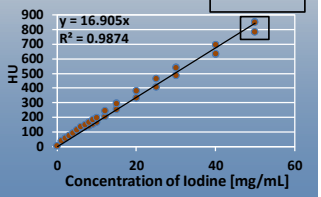




Lack-of-Fit Test

- 2 replicates
- Pure sum of squares
  - Squared sum of distances between data values and mean of data values at i





Lack of Fit Test

- For a good model the Lack of Fit Error will be small compared to the Pure Error
  - Actually compares Mean Square, not Sum Square

Lack-of-Fit Test

- F-Statistic Test
  - H_0: Model is a good fit
  - H_1: Model is a poor fit
- F = 2.7 and p-value = 0.002 → reject H_0

	SumSq	DF	MeanSq	F	pValue
Total	3.9275e+06	72	54548		
Model	3.8827e+06	1	3.8827e+06	6154.2	1.01e-70
Residual	44794	71	630.9		
. Lack of fit	20632	17	1213.6	2.7124	0.0027825
. Pure error	24162	54	447.44		

Lack of Fit Test

- Why is it a poor fit?
  - Data was not actually replicates
  - Different recon kernels, some with artificial HU enhancement
  - Maybe power law is a better model?

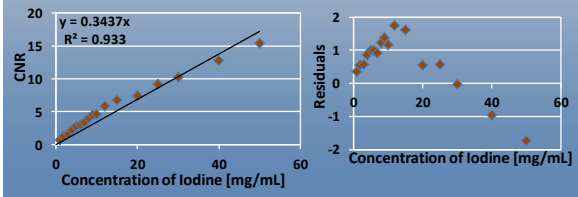
	SumSq	DF	MeanSq	F	pValue
Total	3.9275e+06	72	54548		
Model	3.8827e+06	1	3.8827e+06	6154.2	1.01e-70
Residual	44794	71	630.9		
. Lack of fit	20632	17	1213.6	2.7124	0.0027825
. Pure error	24162	54	447.44		

When can I use SLR?

- LINE
  - Linear
  - Errors are:
    - Independent
    - Normal errors
    - Equal variance errors
- . . . Or all errors are equal, normal, independent, and have a zero mean value

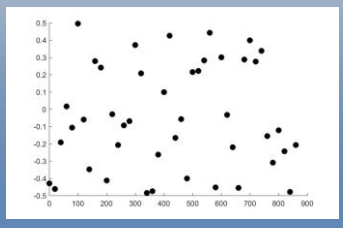
Residuals vs. Fit Plot

- Inspect to determine



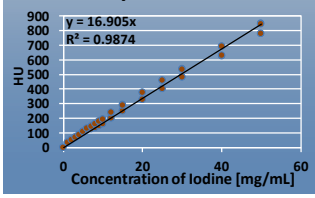
Residuals vs. Fit Plot

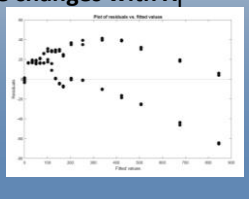
- Ideal: a random, equal sized band about 0



Residuals vs. Fitted Plot

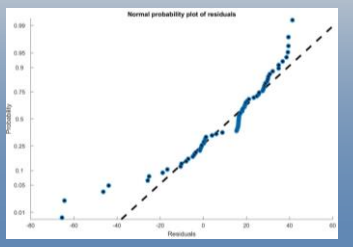
- Structure: Not linear relationship
- Unequal Band: Variance changes with X_i





Testing Normality

- Normal probability plot of residuals
  - If data departs from straight line, then errors are not normal



Pitfall

- The data does NOT have to be normal
  - Errors should be normal
- Normal data typically has normal errors so it satisfies that part of the criteria for use

How strict are the criteria?

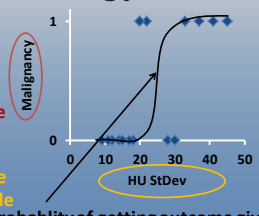
- Equal variance and independence are important
- Prediction intervals are pretty sensitive to normality of errors

Binomial Logistic Regression

- After this you should be able to:
  - Understand basic terminology
  - Interpret results of a regression analysis
  - Determine if a model is "good"
  - Avoid common pitfalls

Terminology

- What we measure
  - Response variable
  - Outcome variable
  - Dependent variable
- What we investigate
  - Predictor variable
  - Explanatory variable
  - Independent variable



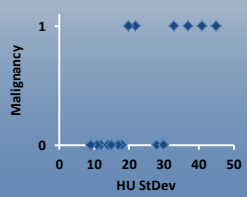
Logistic Regression = Model probability of getting outcome given input

Terminology

- Binary outcome
  - Code outcome as 0 or 1 (e.g., malignancy)
- Ordinal outcome
  - Code outcome as integer 0, 1, etc.
  - Implies an order (e.g., pain scale)
- Nominal outcome
  - Code outcomes, but no natural ordering
  - E.g., basal cell carcinoma, squamous cell carcinoma, and melanoma

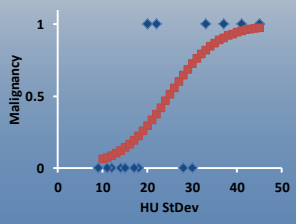
Model Terminology

- Model = P(Malignancy|HU StDev)
  - . . . Let's build up to this
  - More deviation for Malignant tumors
  - $Y_i = W_0 + W_i * X_i$
    - But HU St Dev is not probability!
    - Need something else



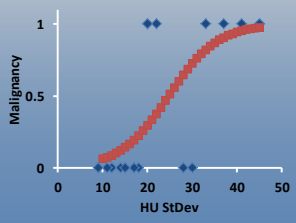
Model Terminology

- Logistic (Logit) function
- $P(M|HU\ St\ Dev) = \frac{1}{1+e^{-a}}$
  - a: $W_0 + W_i * X_i$
    - W_i & W_o: Curve Midpoint & Mid-segment Slope
- Iterative optimization



Interpreting Results

- $W_0 + W_i * X_i$
  - W_0 = -4.571
  - W_i = 0.185
- Rearrange
  $P = \frac{1}{1+e^{-a}}$
- $Ln(P/1+P) = W_0 + W_i * X_i$
- W_i is increase in log odds
  - Clear, right?



Interpreting W_1

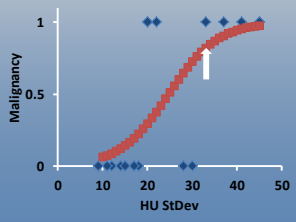
- Log Odds: $Ln(P/1+P) = W_0 + W_1 * X_1$
  - Log Odds: W_1 = 0.185
- Odds Ratio $Exp(W_0) = Exp(0.185) = 1.2$
  - For a unit increase in HU StDev, the odds of malignancy go up 1.2x

Coefficients, Standard Errors, Odds Ratios, and 95% Confidence Limits...

Variable	Coeff.	StdErr	p	O.R.	Low --	High
1	0.1847	0.0818	0.0240	1.2029	1.0247	1.4120
Intercept	-4.5708	1.9425	0.0186			

HU St Dev is 33 what is the probability?

- Ln(P/1+P) = -4.57 + 0.19*(33)
  = 1.524
- Odds Ratio = 4.59
- Probability = OR/1+OR
  =82.1%

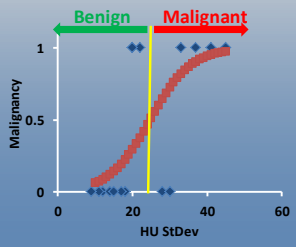


An Extension of the Model

- We only used one predictor (HU StDev)
- Could extend this to more
  - e.g., HU Mean and Tumor Area
    - New Model: $W_0 + W_1*X_1+W_2*X_2+W_3*X_3$
    - W_1*X_1: HU St Dev
    - W_2*X_2: HU Mean
    - W_3*X_3: HU Area

Classification

- Set a threshold to make this a classifier
- Default 50% (0.5)



How good is our model?

- Pseudo-R^2
  - Value is NOT how much variation the logistic model accounts for
  - Several flavors and interpretation is not straight forward
  - Read up if you use this

How good is our model?

- Hosmer-Lemeshow Goodness-of-Fit Test
  - No Replicates Needed
  - Bin St Dev into 5-10 bins of increasing probability of malignancy
    - Want > 5 per bin if possible (We have too little data)
  - Calculate the "actual" probability of malignancy
  - Average the model probabilities for the same point
  - Compare with chi-squared test
    - Gives P-value
  - H_0: Model is a good fit
  - H_1: Model is a poor fit

Hypothesis Testing (Wald Test)

- Is there a significant relationship between HU St Dev and malignancy?
  - Is slope significant?
    - Null (H_0): $W_1=0$
    - Alternate (H_1): $W_1 \neq 0$

Coefficients, Standard Errors, Odds Ratios, and 95% Confidence Limits...

Variable	Coeff.	StdErr	p	O.R.	Low --	High
1	0.1847	0.0818	0.0240	1.2029	1.0247	1.4120
Intercept	-4.5708	1.9425	0.0186			

Statpages.info/logistic

Hypothesis Testing (Wald Test)

- Estimate of $W_1 = 0.1847$
- Standard Error of W_1 Estimate = 0.0818
- Estimate is ***2.26x*** the SE so we get a small p-value and "confirm" HU St Dev adds something to the model
  - Does NOT mean it is the best model

Coefficients, Standard Errors, Odds Ratios, and 95% Confidence Limits...

Variable	Coeff.	StdErr	p	O.R.	Low --	High
1	0.1847	0.0818	0.0240	1.2029	1.0247	1.4120
Intercept	-4.5708	1.9425	0.0186			

Statpages.info/logistic

When can I use Binary Logistic Regression?

- Fewer Assumptions
  - Do NOT need linear relationship between outcome and predictor
  - Do NOT need errors to be normal
  - Do NOT need variance same at each predictor value (Homoscedasticity)

When can I use Binary Logistic Regression?

- Need binary outcome
  - Lose precision if you "make" it binary
    - E.g., Blood pressure: "high" and "low"
- Observations should still be independent
- Predictor must be linearly related to log odds of outcome

Pitfall

- Need relatively large sample size
  - At least 10*# predictors for each outcome
    - For a model with predictors: HU St Dev, HU Mean, and Tumor Area
      - 10*3 = 30
      - So at least 30 malignant and 30 benign
      - Big problem for prospective studies

Pitfall

- Large sample size
  - At least 10*# predictors for each outcome
    - Big problem for prospective studies
    - So if you have 1% incidence in a population, then for a prospective study you would need:
      - (10*3/0.01) = 3,000 participants to get 30 malignancies

Further Reading

- 2016 AAPM Stats Refresher (Handouts and Virtual Library Available)
  - Clinical trials and the medical physicist: design, analysis, and our role
  - Implementation and Analysis of Observer Studies in Medical Physics
  - Analysis of Dependent Variables: Correlation and Simple Regression
  - Hypothesis or Hypotheses, That is the Question
- Penn State Stats 501
  - https://onlinecourses.science.psu.edu/stat501/

Cite This Talk/Handout

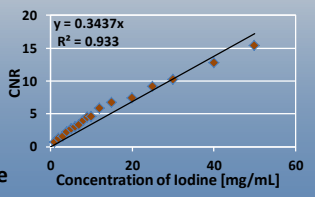
Sensakovic, WF. TU-AB-KDBRA2-3: Linear and logistic regressions - What they try to explain, and how to interpret the results. 60th Annual Meeting of the AAPM, Nashville, TN July 29-August 2, 2018.

Image Copyright Requirements

- https://creativecommons.org/licenses/

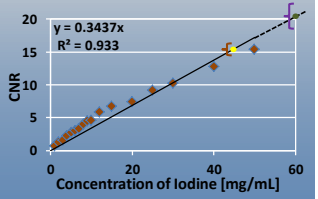
How accurate are SLR model estimated values?

- Confidence Interval
  - New predictor value within our data range
- Prediction Interval
  - New Predictor value outside our data range
  - More uncertainty



How accurate are SLR model estimated values?

- At 45mg/mL, CNR estimate is 15.5
  - 95% CI (14.9, 16.1)
- At 60mg/mL CNR estimate is 20.6
  - 95% PI (18.2, 23.0)



How accurate are BLR model estimated values?

HU St Dev	95% Low	95% High
33.0000	0.3486	0.9752
12.0000	0.0114	0.4377
45.0000	0.4699	0.9995
20.0000	0.0972	0.6165
37.0000	0.3952	0.9930
18.0000	0.0619	0.5563
28.0000	0.2717	0.8991
11.0000	0.0084	0.4239
etc.