SKKU Biostats and Big data

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

Lecture 07

Regression wisdom, re-expressing data

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

Review: Key Points

Chapter 8: Linear Regression

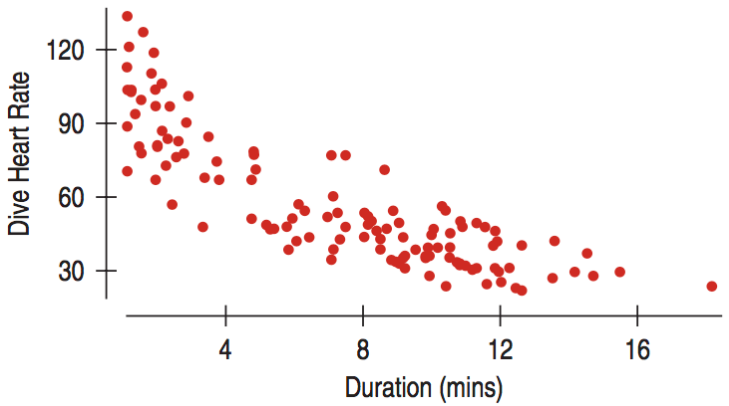
- residual = Observed value (y) – Predicted value ($\hat{y}$)
- Line of "best fit": $\arg\min \sum(y-\hat{y})^2 = \sum d_i^2$: *Least squares* line
- $\hat{y} = b_0 + b_1 x$. Slope, $b_1 = r\frac{s_y}{s_x}$ Intercept, $b_0 = \bar{y} - b_1\bar{x}$
- Residuals $e = y - \hat{y}$
- DATA = MODEL + RESIDUAL: $y = b_0 + b_1 x + e$
- Residual plot should show no interesting pattern.
- $R^2 = 1 - \frac{Sum\ of\ squared\ residuals}{Sum\ of\ squared\ deviation\ from\ the\ mean} = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$. In the linear model, R^2 is same with r^2.
- Predicting y with x and predicting x with y are different!
- Regression to the mean: due to some randomness in the data!

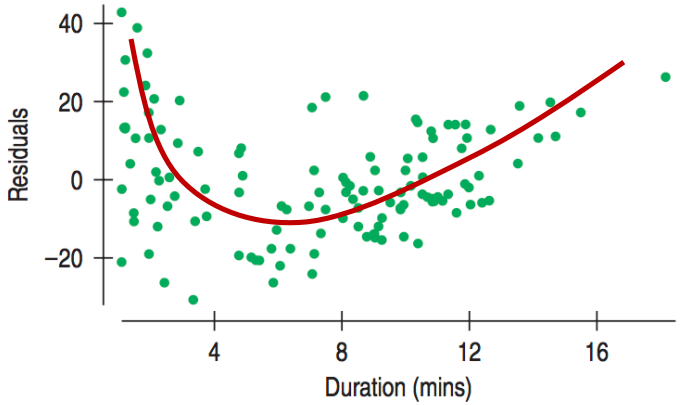
Quiz 07-1

https://forms.gle/JaRna2zq7SqJqWiT7

Regression wisdom

- "Bends" is often not apparent in the scatterplot of raw data. Residual plots will help!





Regression wisdom

- "Bends" is often not apparent in the scatterplot of raw data. Residual plots will help!
- Subgroups? Look at the histogram of the residuals

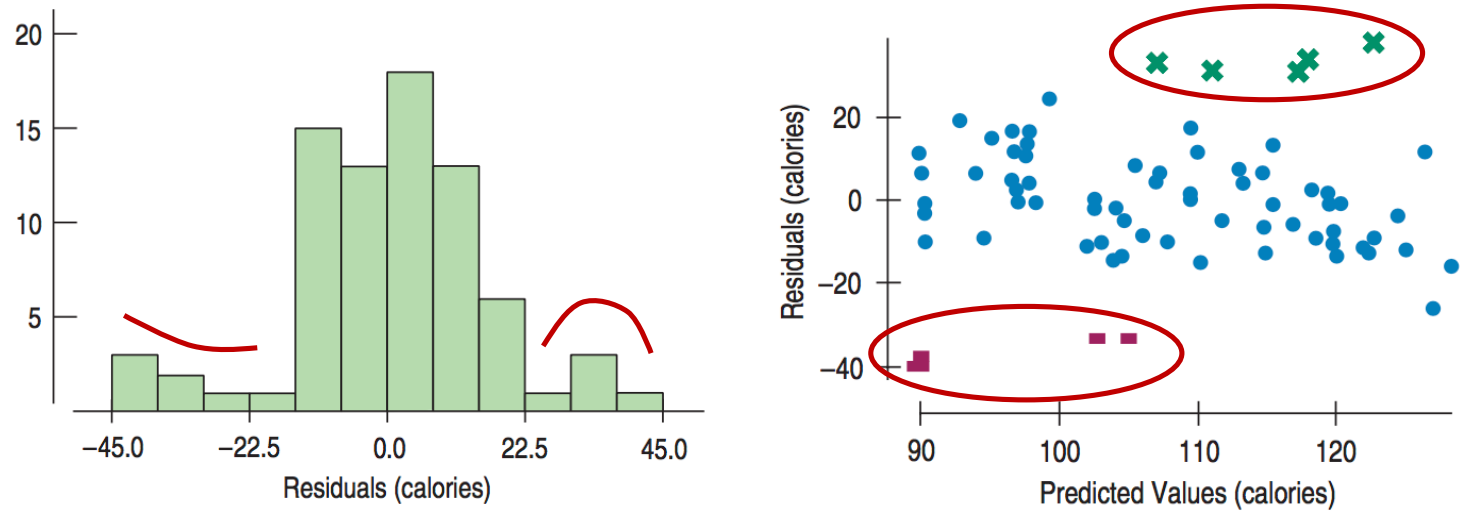
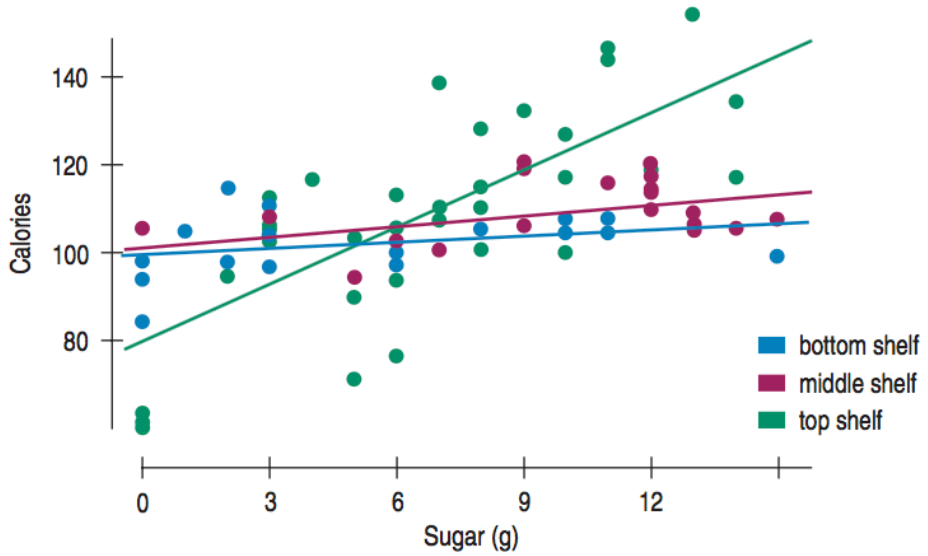


FIGURE 9.3
A histogram of the regression residuals shows small modes both above and below the central large mode. These may be worth a second look.

Regression wisdom

- "Bends" is often not apparent in the scatterplot of raw data. Residual plots will help!
- Subgroups? Look at the histogram of the residuals
- If you already know there are different groups, get the regression lines separately.



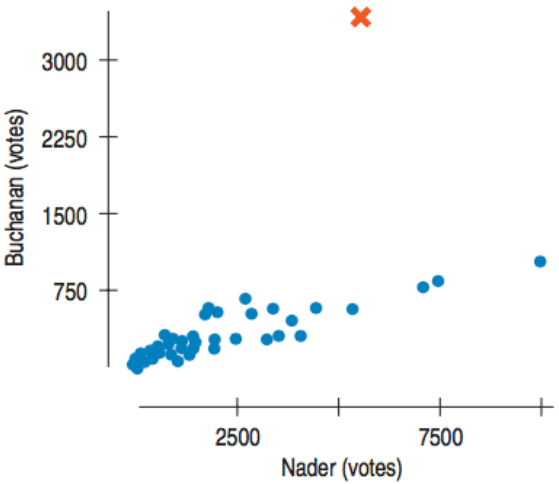
Regression wisdom

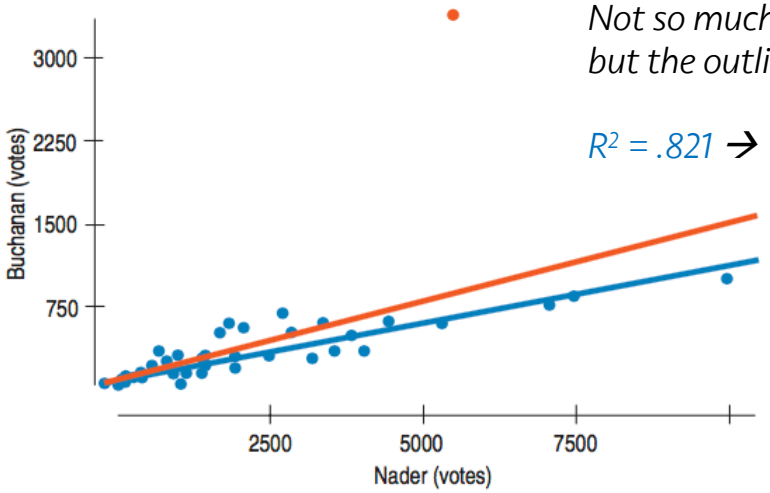
- "Bends" is often not apparent in the scatterplot of raw data. Residual plots will help!
- Subgroups? Look at the histogram of the residuals
- If you already know there are different groups, get the regression lines separately.
- Extrapolation is not warranted (sampling bias, generalization error)
- No causation!
- Outlier!

Regression wisdom

"Give me a place to stand and I will move the Earth."

–Archimedes
(287-211 B.C.E.)





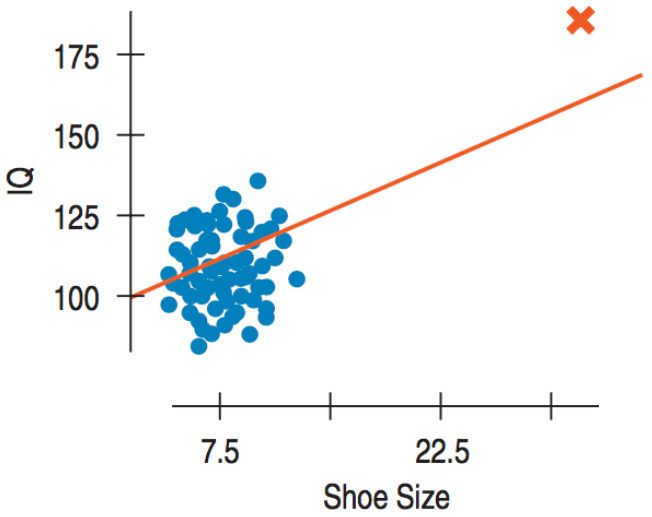
Not so much change in line, but the outlier drops R^2 a lot!

$R^2 = .821$ → $R^2 = .428$

Regression wisdom

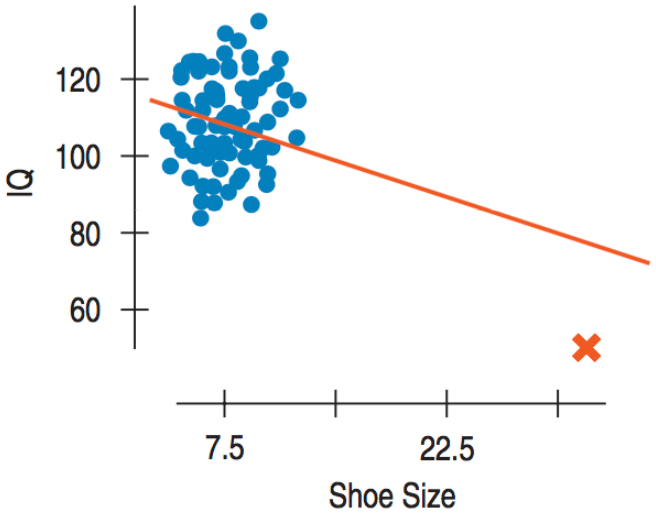
"Give me a place to stand and I will move the Earth."

–Archimedes
(287-211 B.C.E.)



High leverage

changes line a lot
and increases R^2 a lot



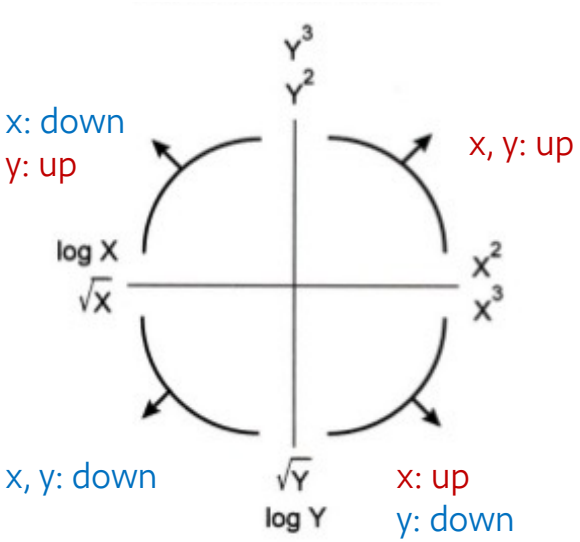
Regression wisdom

- "Bends" is often not apparent in the scatterplot of raw data. Residual plots will help!
- Subgroups? Look at the histogram of the residuals
- If you already know there are different groups, get the regression lines separately.
- Extrapolation is not warranted (sampling bias, generalization error)
- No causation!
- Outlier!
- Lurking variable (again)
- Be cautious in working with summary statistics!

Quiz 07-2

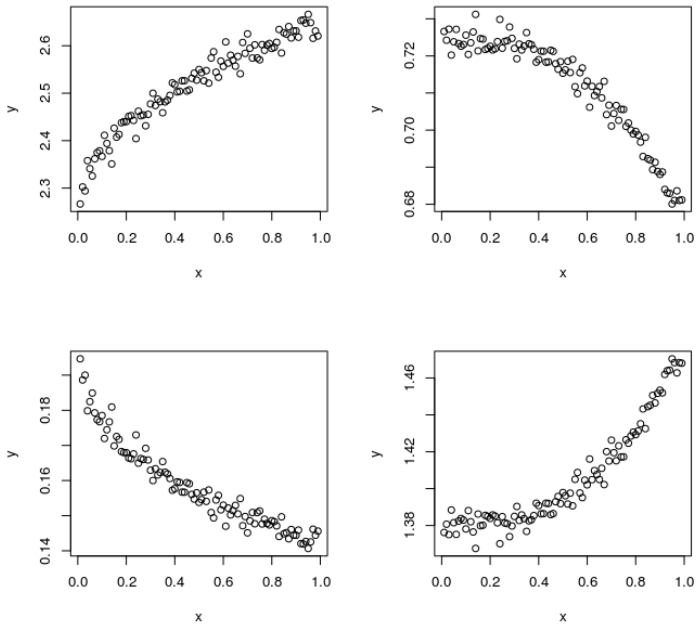
https://forms.gle/ktMASYGyTFxdTvgw7

Re-expressing Data: Tukey's Ladder



Power	Name	Comment
2	The square of the data values, y^2.	Try this for unimodal distributions that are skewed to the left.
1	The raw data—no change at all. This is "home base." The farther you step from here up or down the ladder, the greater the effect.	Data that can take on both positive and negative values with no bounds are less likely to benefit from re-expression.
1/2	The square root of the data values, $\sqrt{y}$.	Counts often benefit from a square root re-expression. For counted data, start here.
"0"	The logarithm of y. It doesn't matter whether you take the log base 10, the natural log, (or any other base).	Measurements that cannot be negative, and especially values that grow by percentage increases such as salaries or populations, often benefit from a log re-expression. When in doubt, start here. If your data have zeros, try adding a small constant to all values before finding the logs.
−1/2	The (negative) reciprocal square root, $-1/\sqrt{y}$.	An uncommon re-expression, but sometimes useful. Changing the sign to take the *negative* of the reciprocal square root preserves the direction of relationships, making things a bit simpler.
−1	The (negative) reciprocal, $-1/y$.	Ratios of two quantities (miles per hour, for example) often benefit from a reciprocal. (You have about a 50–50 chance that the original ratio was taken in the "wrong" order for simple statistical analysis and would benefit from re-expression.) Often, the reciprocal will have simple units (hours per mile). Change the sign if you want to preserve the direction of relationships. If your data have zeros, try adding a small constant to all values before finding the reciprocal.

Re-expressing Data: Tukey's Ladder



Power	Name	Comment
2	The square of the data values, y^2.	Try this for unimodal distributions that are skewed to the left.
1	The raw data—no change at all. This is "home base." The farther you step from here up or down the ladder, the greater the effect.	Data that can take on both positive and negative values with no bounds are less likely to benefit from re-expression.
1/2	The square root of the data values, $\sqrt{y}$.	Counts often benefit from a square root re-expression. For counted data, start here.
"0"	The logarithm of y. It doesn't matter whether you take the log base 10, the natural log, (or any other base).	Measurements that cannot be negative, and especially values that grow by percentage increases such as salaries or populations, often benefit from a log re-expression. When in doubt, start here. If your data have zeros, try adding a small constant to all values before finding the logs.
−1/2	The (negative) reciprocal square root, $-1/\sqrt{y}$.	An uncommon re-expression, but sometimes useful. Changing the sign to take the *negative* of the reciprocal square root preserves the direction of relationships, making things a bit simpler.
−1	The (negative) reciprocal, $-1/y$.	Ratios of two quantities (miles per hour, for example) often benefit from a reciprocal. (You have about a 50–50 chance that the original ratio was taken in the "wrong" order for simple statistical analysis and would benefit from re-expression.) Often, the reciprocal will have simple units (hours per mile). Change the sign if you want to preserve the direction of relationships. If your data have zeros, try adding a small constant to all values before finding the reciprocal.

Quiz 07-3

https://forms.gle/3FunTwueHLcEmob19

Re-expressing Data: Logarithms

- When none of the data values is zero or negative:

Model Name	*x*-axis	*y*-axis	Comment
Exponential	x	$\log(y)$	This model is the "0" power in the ladder approach, useful for values that grow by percentage increases.
Logarithmic	$\log(x)$	y	A wide range of x-values, or a scatterplot descending rapidly at the left but leveling off toward the right, may benefit from trying this model.
Power	$\log(x)$	$\log(y)$	The Goldilocks model: When one of the ladder's powers is too big and the next is too small, this one may be just right.

- Watch out for negative data values
  - ➢ one possible cure for zeros and small negative values is to add a constant.
- Watch for data far from 1
  - ➢ may not be much affected by re-expression
  - ➢ consider subtracting a constant to bring them back near 1 (e.g., years, 2000 -> 1)

Key Points

Chapter 9 and 10: Regression wisdom, re-expressing data

- Bends, subgroups, outliers
- Cautious about extrapolation, causation, lurking variables, summary stats
- Tukey's ladder