MSCA 31000 - Introduction to Statistical Concepts

Chapter 14

Q2: The formula for a regression equation based on a sample size of 25 observations is Y' = 2X + 9. (a) What would be the predicted score for a person scoring 6 on X? (b) If someone's predicted score was 14, what was this person's score on X?

Answer:

a) 21
b) 2.5

Q4: What does the standard error of the estimate measure? What is the formula for the standard error of the estimate?

Answer: The standard error of the estimate is a measure of the accuracy of predictions. The formula is

$$\sigma_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N}}$$

where,

- σ_{est} is the standard error of the estimate
- Y is an actual score
- Y' is a predicted score
- N is the number of pairs of scores

Q6: For the X,Y data below, compute:

(a) r and determine if it is significantly different from zero.
(b) the slope of the regression line and test if it differs significantly from zero.
(c) the 95% confidence interval for the slope.

Answer:

a) r = 0.91. It is significantly different from zero.
b) Slope b = r * s_x / s_y = .91 * 1.48 / 3.11 = 0.433
c) t = statistic - hypothesized value / estimated std error of statistic

   s_{est} = 3.6
   SSX = 8.8
   s_b = 1.21

t = b / s_b = 0.433 / 1.21 = 0.35785
df = N - 2 = 3

2*pt(0.35785,3,lower.tail=F) = .744. Slope is not significantly different from 0.

95% Confidence interval for slope b = (b - t * s_b, b + t * s_b) = **(3.42, 4.28)**

Q8: The correlation between years of education and salary in a sample of 20 people from a certain company is .4. Is this correlation statistically significant at the .05 level?

Answer:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

r = 0.4
n = 20
df = n - 2 = 18
t = 1.85164

pt(1.85164,20,lower.tail=F) = **.039**

Therefore, the correlation **is statistically significant** at the .05 level.

Q10: Using linear regression, find the predicted post-test score for someone with a score of 43 on the pre-test.

Answer: lm(Post ~ Pre, data=q10), where q10 is the dataset
Intercept = A = 16.1552
Slope = b = 0.7869

Using formula, Y = b * X + A
where,
Y = Post score
X = Pre score
b = slope
A = intercept

$Y_{(X=43)}$ = .7869 * 43 + 16.1552 = **49.99**

Q12: Based on the table below, compute the regression line that predicts Y from X.

Answer: Equation of regression like is Y’ = bX + A

where,

- b = r * Sy/Sx = -0.6 * 3/2.5 = -0.72
- A = My - b*Mx = 12 - (-.72 * 10) = 19.2

Therefore, the regression line equation is **Y' = -0.72X + 19.2**

Q14: True/false: If the slope of a simple linear regression line is statistically significant, then the correlation will also always be significant.

Answer: True

Q16: True/false: If the correlation is .8, then 40% of the variance is explained.

Answer: False. $r^2 = 0.64$, so 64% of the variance is explained.

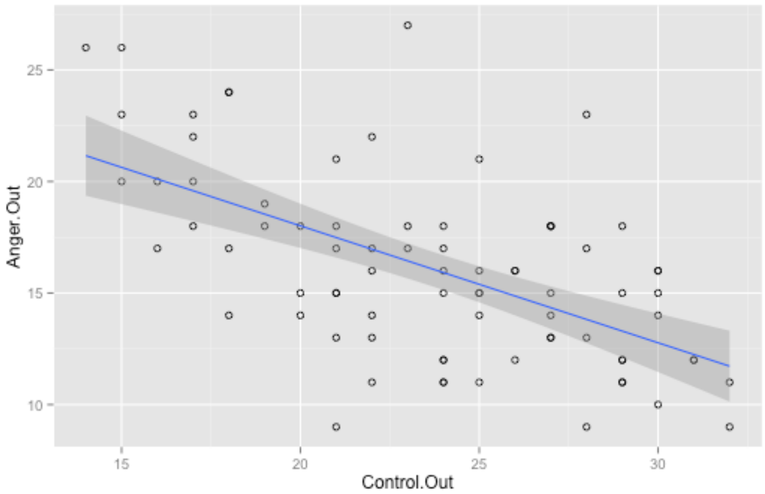
Q18: (AM#23) Find the regression line for predicting Anger-Out from Control-Out.

(a) What is the slope?
(b) What is the intercept?
(c) Is the relationship at least approximately linear?
(d) Test to see if the slope is significantly different from 0.
(e) What is the standard error of the estimate?

Answer:

```
My <- mean(angerdata$Anger.Out)
Mx <- mean(angerdata$Control.Out)
sy <- sd(angerdata$Anger.Out)
sx <- sd(angerdata$Control.Out)
r <- cor(angerdata$Control.Out,angerdata$Anger.Out)
```

(a) slope = b = r * sy / sx = -0.524132
(b) intercept = A = My - b*Mx = 28.49482
(c) From the plot (and the slope value), we can see the relationship is approximately linear.



(d) t = (statistic - hypothesized value) / estimated std error of statistic
I consider this a 1-sample t-test since the Anger.Out and Control.Out variables are referring to the same person.

r = -0.5826834
s_{est} = 3.45
SSX = 21.98202
s_b = 0.7358432
df = 78 - 2 = 76

t = b / s_b = -0.7122876

p-value of two tailed test = 2*pt(-0.7122876,76) = **0.4784677**. Therefore, slope is not significantly different from 0.

(e) s_{est} = 3.45

Chapter 16

Q1. When is a log transformation valuable?

Answer: A log transformation is useful to reduce skew in the distribution, typically a positive skew (ref: http://fmwww.bc.edu/repec/bocode/t/transint.html)

Q2: If the arithmetic mean of log10 transformed data were 3, what would be the geometric mean?

Answer: Geometric mean = 10$^{\wedge AM}$ = 10^3 = 1000

Q3: Using Tukey's ladder of transformation, transform the following data using a λ of 0.5: 9, 16, 25

Answer: λ of 0.5 implies square root.

- sqrt(9) = 3
- sqrt(16) = 4
- sqrt(25) = 5

Q4: What value of λ in Tukey's ladder decreases skew the most? Q5: What value of λ in Tukey's ladder increases skew the most?

For Q4 and 5, I am sharing the answer I gave on the class discussion board. Prof Barlas' explanations differ from my answer, but I am answering based on some code experiments I ran in R.

Summary: **I believe factors such as the distribution of the data, whether the data values are normalized and the kurtosis will all contribute in varying degrees towards deciding an appropriate lambda.**

Detailed:

Conceptually, the point about shrinking the extremities using square root or smoothing the distribution by squaring makes sense. I tried to verify that square root shrinks the extreme observations the most and thus reduces the skew the most (R code attached) for both normal and exponential distributions, but I don't see a meaningful transformation.

Also, according to the link you shared (http://fmwww.bc.edu/repec/bocode/t/transint.html), the description for square root says:

"Square root, x to x^(1/2) = sqrt(x), is a transformation with a moderate effect on distribution shape: it is weaker than the logarithm and the cube root. It is also used for reducing right skewness, and also has the advantage that it can be applied to zero values. Note that the square root of an area has the units of a length. It is commonly applied to counted data, especially if the values are mostly rather small."

The highlighted section indicates there are caveats to when a square root is appropriate for a transformation.

Given this, **I believe factors such as the distribution of the data, whether the data values are normalized and the kurtosis will all contribute in varying degrees towards deciding an appropriate lambda.** Do you agree to this?