Pol Sci 630: Problem Set 3 - Comparisons and Inference - Solutions

Prepared by: Jan Vogler (jan.vogler@duke.edu)

Due Date: Tuesday, September 15th, 2015, 10 AM (Beginning of Class)

R Programming

Problem 1

```
### Problem 1:

### a

x = seq(1, 1000, by = 1)
y = 2 * x - 5

cov(x, y)

## [1] 166833.3
```

Interpretation: the covariance indicates that there might be a positive linear relationship of x and y. However, because the covariance is not to-scale, the number itself is not very meaningful.

Note: if someone created a different kind of linear function, the result might be a covariance that indicates a negative linear relationship.

```
cor(x, y)

## [1] 1
```

Interpretation: the correlation is bound between -1 and 1. The correlation value of 1 here means that x and y have a perfect positive linear relationship. As x goes above its mean, y goes above its mean. This is not surprising as we created y through a linear function of x.

Note: if someone created a different kind of linear function, the result might be a correlation value of -1, indicating a perfect negative linear relationship. As x goes above its mean, y goes below its mean.

```
noise = rnorm(1000, mean = 0, sd = 10)
y2 = y + noise

cov(x, y2)

## [1] 166692
```

Interpretation: same as above - the covariance indicates that there might be a linear positive relationship of x and y. However, because the covariance is not to-scale, the number itself is not very meaningful.

```
cor(x, y2)

## [1] 0.9998528
```

The correlation is bound between -1 and 1. The correlation value here should be very close to 1 or -1 but not be exactly that value due to the random error. As long as some random error has been introduced to a formerly perfect linear relationship, even if that relationship is still generally linear, there will be a reduction in the absolute value of the correlation. Accordingly, the result you can expect to get here is an absolute value of approximately 0.99. The exact value, however, depends on the size of the random error that you introduced. If you introduce a random error that has a greater variance, then the value of your correlation will go down further.

Interpretation: a correlation close to 1 or -1 indicates a nearly perfect linear relationship of two variables. If the relationship is positive, the following is true: as x goes above its mean, y goes above its mean. If the relationship is negative, the following is true: as x goes above its mean, y goes below its mean. The randomly distributed error will in most cases not change the generally strong linear relationship between the two variables.

```r
### b

correlation = function(v1, v2) {
    numerator = sum((v1 - mean(v1)) * (v2 - mean(v2)))/(length(a) - 1)
    denominator = sd(v1) * sd(v2)
    print(numerator/denominator)
}

### Let's try this function.

a = seq(1, 10, by = 1)
noise2 = rnorm(10, mean = 0, sd = 1)
b = a + noise2

cor(a, b)
```

```
## [1] 0.9355679
```

```r
correlation(a, b)
```

```
## [1] 0.9355679
```

```r
# These two return the same result, meaning that we did it correctly.

### c

correlation2 = function(v1, v2) {
    if (length(v1) == length(v2)) {
        if (is.numeric(v1) & is.numeric(v2)) {
            numerator = sum((v1 - mean(v1)) * (v2 - mean(v2)))/(length(a) -
                1)
            denominator = sd(v1) * sd(v2)
            print(numerator/denominator)
        } else {
            print("The two vectors need to be numeric.")
        }
```

```
    } else {
        print("The two vectors need to be of the same length.")
    }
}

### Let's plug in vectors that do not work.

c = seq(1, 11)
length(c)

## [1] 11

correlation2(a, c)

## [1] "The two vectors need to be of the same length."

# Returns the correct error message.

d = c("a", "b", "c", "d", "e", "f", "g", "h", "i", "j")
is.numeric(d)

## [1] FALSE

correlation2(a, d)

## [1] "The two vectors need to be numeric."

# Returns the correct error message.
```

Problem 2

```
### Problem 2

### a
```

```r
data(swiss)
summary(swiss)
```

```
##    Fertility      Agriculture     Examination      Education
##  Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00
##  1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00
##  Median :70.40   Median :54.10   Median :16.00   Median : 8.00
##  Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98
##  3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00
##  Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00
##     Catholic       Infant.Mortality
##  Min.   :  2.150   Min.   :10.80
##  1st Qu.:  5.195   1st Qu.:18.15
##  Median : 15.140   Median :20.00
##  Mean   : 41.144   Mean   :19.94
##  3rd Qu.: 93.125   3rd Qu.:21.70
##  Max.   :100.000   Max.   :26.60
```

```r
### b
```

```r
lm1 = lm(Education ~ Fertility + Agriculture + Examination + Catholic + Infant.Mortality
    data = swiss)
```

```r
summary(lm1)
```

```
##
## Call:
## lm(formula = Education ~ Fertility + Agriculture + Examination +
##     Catholic + Infant.Mortality, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3949  -2.3716  -0.2856   2.8108  11.2985
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      32.74414    8.87888   3.688 0.000657 ***
## Fertility        -0.40851    0.08585  -4.758 2.43e-05 ***
## Agriculture      -0.16242    0.04488  -3.619 0.000804 ***
## Examination       0.41980    0.16339   2.569 0.013922 *
## Catholic          0.10023    0.02150   4.663 3.29e-05 ***
## Infant.Mortality  0.20408    0.28390   0.719 0.476305
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.907 on 41 degrees of freedom
## Multiple R-squared:  0.7678,Adjusted R-squared:  0.7395
## F-statistic: 27.12 on 5 and 41 DF,  p-value: 5.223e-12
```

c) In order to get full points on this problem, you need an interpretation for each of the 5 variables.

The interpretation would look like this for Fertility:

There is a negative linear relationship between Fertility and Education. For a 1-point increase in Fertility, we expect a 0.41-point decrease in Education, holding all other variables constant. The t-value is -4.758. This t-value implies a p-value of $2.43 * 10^{-5}$. This $p < 0.001$ corresponds to a type-1 error rate of $\alpha < 0.001$, meaning that the statistical relationship is significant at all common levels of statistical significance.

The other variables are interpreted accordingly. Agriculture and Catholic are significant at all common levels of statistical significance as well. Please note that Examination is significant at a level of $p < 0.05$, $\alpha < 0.05$, and Infant.Mortality is not significant at common levels of statistical significance. The levels of significance can be found in the tutorial notes.

What can we say about causality? Nothing really. There are two primary reasons for this:

First and foremost, linear regression does not per se tell us anything about causality - it primarily measures correlation between variables.

Second, we do not have any theory regarding the relationship of Education on the other covariates and so we cannot make any causal claims that are grounded in theory. In particular, there might be a mutual influence between Education and the other variables that we

regress it on. This phenomenon is called "endogeneity" and there are various ways to deal with it that you will learn about in the class.

In short, we can't say anything about causality here.

Covariance and Correlation Mathematically

Problem 3

a) Both X and Y have the uniform distribution over all four points, so each outcome is equally likely.

$\mathbb{E}(X) = \frac{1}{4} * (-1) + \frac{1}{4} * (0) + \frac{1}{4} * (0) + \frac{1}{4} * (1) = 0$

$\mathbb{E}(Y) = \frac{1}{4} * (0) + \frac{1}{4} * (1) + \frac{1}{4} * (-1) + \frac{1}{4} * (0) = 0$

$\mathbb{E}(X * Y) = \frac{1}{4} * (0) + \frac{1}{4} * (0) + \frac{1}{4} * (0) + \frac{1}{4} * (0) = 0$

$Cov(X, Y) = \mathbb{E}(X * Y) - \mathbb{E}(X) * \mathbb{E}(Y) = 0 - 0 * 0 = 0$

Proof by contradiction. In order to prove that X and Y are not independent, it is sufficient to show that they violate the necessary conditions for independence in one case.

For independence, the following must be true:

$Pr(X = x \cap Y = y) = Pr(X = x) * Pr(Y = y)$

Given this definition, it is sufficient to show that this is not true for some values of X. For example:

$Pr(X = -1 \cap Y = 1) = 0$ because this event never occurs.

$Pr(X = -1) = \frac{1}{4}$ and $Pr(Y = 1) = \frac{1}{4}$, meaning that $Pr(X = -1) * Pr(Y = 1) = \frac{1}{16}$

Accordingly, $Pr(X = -1) * Pr(Y = 1) = \frac{1}{16} \neq 0 = Pr(X = -1 \cap Y = 1)$

b) $\mathbb{E}(X) = \frac{1}{3} * (-1) + \frac{1}{3} * (0) + \frac{1}{3} * (1) = 0$

$\mathbb{E}(Y) = \frac{1}{3} * (1) + \frac{1}{3} * (0) + \frac{1}{3} * (1) = \frac{2}{3}$

$\mathbb{E}(X * Y) = \frac{1}{3} * (-1) + \frac{1}{3} * (0) + \frac{1}{3} * (1) = 0$

$Cov(X, Y) = \mathbb{E}(X * Y) - \mathbb{E}(X) * \mathbb{E}(Y) = 0 - \frac{2}{3} * 0 = 0$

Are X and Y independent? No. We don't need to prove this because we know that $Y = X^2$, so Y was defined to be a function of X. The reason why we don't capture their dependence is that they are not *linearly dependent*. Instead, they are dependent through a quadratic function.

c) In order to solve this problem, as in the above problems, we need to calculate the following:

$Cov(X, Y) = \mathbb{E}(X * Y) - \mathbb{E}(X) * \mathbb{E}(Y)$

In order to do this, R is extremely helpful.

```
### Problem 3

### c

### In order to calculate E(XY), use the following:

sum = 0
for (i in 1:6) {
    for (j in 1:6) {
        sum = (i + j) * (i - j) + sum
    }
}
```

This calculation returns the sum of all 36 outcomes, for all possible combinations of the first and the second dice. Each of the above outcomes is equally likely with probability 1/36. We can either multiply every single value by 1/36 or we can, alternatively, simply divide the sum by 36 to get of $E(X * Y)$. The same applies to $E(X)$ and $E(Y)$ below.

```
exy = sum/36  # = 0
exy

## [1] 0

### In order to calculate E(X), use the following:

sum2 = 0
```

```
for (i in 1:6) {
    for (j in 1:6) {
        sum2 = (i + j) + sum2
    }
}

ex = sum2/36
ex

## [1] 7

### In order to calculate E(Y), use the following:

sum3 = 0
for (i in 1:6) {
    for (j in 1:6) {
        sum3 = (i - j) + sum3
    }
}

ey = sum3/36
ey

## [1] 0

### The covariance is given by the following formula:

exy - ex * ey  # Returns 0.

## [1] 0
```

Why aren't X and Y independent? Let's try a similar proof by contradiction like above. $Pr(X = 12 \cap Y = 1) = 0$ because this event never occurs.

$Pr(X = 12) = \frac{1}{36}$ and $Pr(Y = 1) = \frac{4}{36} = \frac{1}{9}$, meaning that $Pr(X = 12) * Pr(Y = 1) = \frac{1}{324}$

Accordingly, $Pr(X = 12) * Pr(Y = 1) = \frac{1}{324} \neq 0 = Pr(X = 12 \cap Y = 1)$