679 Assignment 1

Dave Anderson

January 31, 2019

3.1, 3.2, 3.5, 3.6, ,3.11, 3.12, 3.13, 3.14

3.1)

The coefficients show the relationship that three variables (TV, radio, and newspaper advertising) have with sales. The null hypothesis for each is that there is no relationship between that individual variable and the response variable of sales, or, in other words, that the variable's corresponding beta coefficient is zero. For TV and Radio, the p value is very small. This indicates that the relationship between each of these variables and sales is not likely to appear just by chance. Newspaper advertising's coefficient has a very high standard error compared to the coefficient, leading to a large p-value, which tells us that newspaper advertising has a weak relationship with sales.

3.2)

Both KNN Classification and KNN Regression identify a neighborhood of the sample space closely related to our x_0 prediction. KNN is typically used for classification problems and calculates a probability that our prediction point falls within a certain class. Regression is utilized for quantitative responses and creates a function to represent the neighborhood and make a prediction.

3.5)

3.6)

Our linear regression takes the form:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

By definition:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Therefore if we use that and replace x with $\bar{x}$ we can conclude:

$$y = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x}$$

$$y = \bar{y}$$

3.11)

a)

```
set.seed(1)
x=rnorm(100)
y=2*x+rnorm (100)

lm1 <- lm(y ~ x + 0)
summary(lm1)
```

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9154 -0.6472 -0.1771  0.5056  2.3109
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x   1.9939     0.1065   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9586 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

Our coefficient estimate is very close to two, which we should expect because a coefficient of two was used to generate our x and y values. The standard error and p values are both very small, which indicates the relationship of y being twice as large as x is not occuring by chance alone.

b)

```
lm2 <- lm(x ~ y +0)
summary(lm2)
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8699 -0.2368  0.1030  0.2858  0.8938
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## y  0.39111    0.02089   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 99 degrees of freedom
```

```
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

Again we have a small p value and standard error, allowing us to reject the null hypothesis that there is no relationship between x and y. According to our coefficient, each unit increase in y leads to a .4 increase in x

c)

Both models had the same t value and very small p values

d)

Numerically:

```r
n <- length(x)
t <- sqrt(n - 1)*(x %*% y)/sqrt(sum(x^2) * sum(y^2) - (x %*% y)^2)
t
```

```
##          [,1]
## [1,] 18.72593
```

e)

$x_i y_i$ and $x_j y_j$ are always being multiplied in our formula. if you replace x_i with y_i, the formula would produce the same results because of associative property.

f)

```r
lm3 <- lm(y ~ x)
summary(lm3)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8768 -0.6138 -0.1395  0.5394  2.3462
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03769    0.09699  -0.389    0.698
## x            1.99894    0.10773  18.556   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9628 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
lm4 <- lm(x ~ y)
summary(lm4)
```

```
##
## Call:
## lm(formula = x ~ y)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90848 -0.28101  0.06274  0.24570  0.85736
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.03880    0.04266    0.91    0.365
## y            0.38942    0.02099   18.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4249 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

3.12)

a)

$$\hat{\beta} = \sum_i x_i y_i / \sum_j x_j^2$$

for regression of y onto x. With regression of x onto y, the only real difference is the x changes to a y on the bottom of the fraction. Therefore, the coefficients are the same if and only if:

$\sum_j x_j^2 = \sum_j y_j^2$

b)

```r
set.seed(1)

x <- 1:100

y <- 3 * x + rnorm(100)

lmx <- lm(y ~ x + 0)
lmy <- lm(x ~ y + 0)

summary(lmx)
```

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.23590 -0.62560  0.04426  0.58507  2.30926
##
## Coefficients:
```

```
##   Estimate Std. Error t value Pr(>|t|)
## x 3.001514   0.001548    1939   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9005 on 99 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.759e+06 on 1 and 99 DF,  p-value: < 2.2e-16
```

```r
summary(lmy)
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76774 -0.19401 -0.01353  0.20963  0.74527
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## y 0.3331564  0.0001718    1939   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3 on 99 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.759e+06 on 1 and 99 DF,  p-value: < 2.2e-16
```

c)

```r
x <- 1:100
y <- 1:100

lmx2 <- lm(y ~ x + 0)
lmy2 <- lm(x ~ y + 0)

summary(lmx2)
```

```
## Warning in summary.lm(lmx2): essentially perfect fit: summary may be
## unreliable
```

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.082e-13 -2.094e-15  2.900e-17  2.218e-15  1.294e-14
##
## Coefficients:
##    Estimate Std. Error   t value Pr(>|t|)
## x 1.000e+00  5.379e-17 1.859e+16   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.129e-14 on 99 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.457e+32 on 1 and 99 DF,  p-value: < 2.2e-16
```

```
summary(lmy2)
```

```
## Warning in summary.lm(lmy2): essentially perfect fit: summary may be
## unreliable
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.082e-13 -2.094e-15  2.900e-17  2.218e-15  1.294e-14
##
## Coefficients:
##    Estimate Std. Error   t value Pr(>|t|)
## y 1.000e+00  5.379e-17 1.859e+16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.129e-14 on 99 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.457e+32 on 1 and 99 DF,  p-value: < 2.2e-16
```

13)

a)

```
set.seed(1)
x <- rnorm(100)
```

b)

```
eps <- rnorm(100, sd = sqrt(0.25))
```

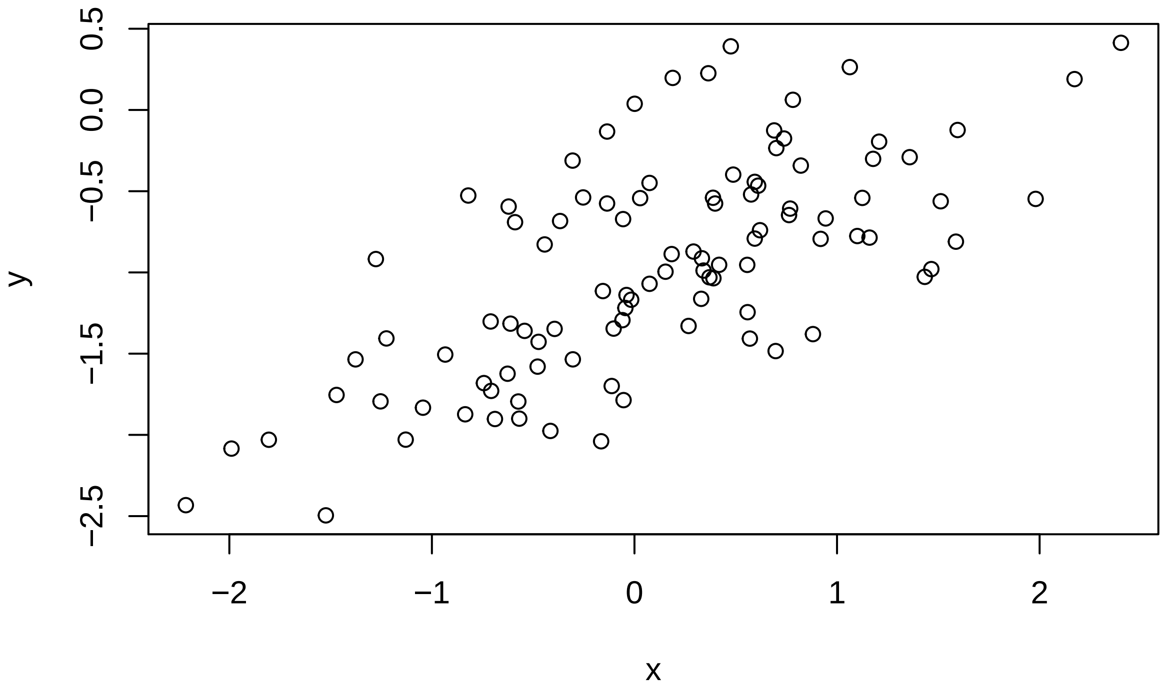
c)

```
y <- -1 + 0.5 * x + eps
```

The length of y is 100, same as x, and $\beta_0 = -1 \beta_1 = 0.5$

d)

```
plot(x,y)
```



The relationship between x and y is linear, as we would expect. It is not a perfect fit because of our generated noise from adding the error. The residuals would be close to normally distributed, since our added error was created from a normal distribution