STAT 231 — Linear Models — Fall 2021

Homework 4 – Simple Linear Regression

NAME: Andrew Guo

Directions: Do all of your work on these sheets and staple them together before you hand them in. You must show all of your work to receive full credit. Put a $\boxed{\text{box}}$ around your final answer, whenever possible.

1. (10 pts) Data from a sample of 10 pharmacies are used to examine the relation between prescription sales volume and the percentage of prescription ingredients purchased directly from the supplier.

x^2	$x_i y_i$	Pharmacy	Ingredients Purchased Directly x (in %)	Sales Volume y (in $1,000)
100	10(25) = 250	1	10	25
324	18(55) = 950	2	18	55
625	1250	3	25	50
1600	3600	4	40	75
2500	5500	5	50	110
3969	8694	6	63	138
1784	3780	7	42	90
900	1800	8	30	60
25	50	9	5	10
3025	5500	10	55	100
$\boxed{14832}$	$\boxed{30814}$	← totals		

$\bar{x} = \dfrac{10+18+\dots+55}{10} \approx 33.8$

$\bar{y} = \dfrac{25+55+\dots+100}{10} \approx 71.3$

$n = 10$

$\bar{x}(\bar{y}) \approx 2409.94$

(a) Find the least-squares estimates for the slope and intercept of the regression line $y = \beta_0 + \beta_1 x + \epsilon$. **DO THIS CALCULATION BY HAND. Hint: make a table containing the relevant information.**

(b) Predict sales volume for a pharmacy that purchases 15% of its prescription ingredients directly from the supplier.

(c) Interpret the value of β_1 in the context of the problem.

a.) $\hat{\beta}_1 = \dfrac{\sum x_i y_i - n(\bar{x})(\bar{y})}{\sum(x_i^2) - n(\bar{x}^2)} = \dfrac{30814 - 10(2409.94)}{14832 - 10(33.8^2)} = \dfrac{6714.6}{3407} \approx 1.97$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1(\bar{x}) = 71.3 - 1.97(33.8) = 4.714$

b.) $y = 4.714 + 1.97x$

$f(15) = 4.714 + 1.97(15) = \boxed{34.264}$ (in $1000)

c.) A β_1 of ≈ 1.97 means that for every increase by 1% in the amount of ingredients purchased by the pharmacy, the sales volume (in $1,000) increases by a factor of 1.97 units.

2. (10 pts) Consider a study which related the crime rate in a major city to the number of casino employees in that city. The study was attempting to associate an increase in crime rate with increasing levels of casino gambling which is reflected in the number of people employed in the gaming industry.

Year	Number of Casino Employees, x (thousands)	Crime Rate, y (per 1,000 population)
1994	20	1.32
1995	23	1.67
1996	29	2.17
1997	27	2.70
1998	30	2.75
1999	34	2.87
2000	35	3.65
2001	37	2.86
2002	40	3.61
2003	43	4.25

$n = 10$

$\bar{x} = \frac{20+23+\dots+43}{10} \approx 31.8$

$\bar{y} = 2.785$

$\bar{x}^2 = 1011.24$

$\bar{x}(\bar{y}) = 88.563$

(a) Find the least-squares estimates for the slope and intercept of the regression line $y = \beta_0 + \beta_1 x + \epsilon$. **DO THIS CALCULATION BY HAND. Hint: make a table containing the relevant information.**

(b) Interpret the value of β_1 in the context of the problem.

(c) Can one interpret the value of β_0 in this problem? Explain.

a.)

x	y	$x_i y_i$	x_i^2
20	1.32	26.4	400
23	1.67	38.41	529
29	2.17	62.93	841
27	2.7	72.9	729
30	2.75	82.5	900
34	2.87	97.58	1156
35	3.65	127.75	1225
37	2.86	105.82	1369
40	3.61	144.4	1600
43	4.25	182.75	1849

$\sum x_i y_i = 941.44$

$\sum x_i^2 = 10598$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n(\bar{x})(\bar{y})}{\sum(x_i^2) - n(\bar{x}^2)} = \frac{941.44 - 10(88.563)}{10598 - 10(1011.24)}$$

$$= \frac{941.44 - 885.63}{10598 - 10112.4} = \frac{55.81}{485.6} \approx 0.115$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1(\bar{x}) = 2.785 - 0.115(31.8) = -0.872!$$

b) For every increase in 1000 casino employees, the crime rate (per 1,000 people) in a major city increases by a factor of 0.115 units.

c.) It doesn't make sense to interpret the value of β_0 here. This would literally mean that if there are no casino employees, the crime rate would be negative. This is impossible; everyone commits sin. This probably indicates that we need to consider other variables that could alter the crime rate; only using the number of casino employees to explain the crime rate could lead to that variable overestimating its effect on the dependent variable.

Hwk 4-3

Andrew Guo

10/7/2021

3.)

(a) Read in the data using the read.csv function

```
pharmaLSRL = read.csv("Hwk4-3.csv")
```

(b) Fit a linear model using the lm function with "Ingredients Purchased Directly" as the predictor and "Sales Volume" as the response. Use the summary function to output the resulting fit

```
attach(pharmaLSRL)

pharmaLSRL.fit <- lm(Sales.Volume~Ingredients.Purchased.Directly, data = pharmaLSRL)

summary(pharmaLSRL.fit)
```

```
##
## Call:
## lm(formula = Sales.Volume ~ Ingredients.Purchased.Directly, data = pharmaLSRL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.074  -4.403  -1.607   5.719  14.834
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      4.6979     5.9520   0.789    0.453
## Ingredients.Purchased.Directly   1.9705     0.1545  12.750 1.35e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.022 on 8 degrees of freedom
## Multiple R-squared:  0.9531, Adjusted R-squared:  0.9472
## F-statistic: 162.6 on 1 and 8 DF,  p-value: 1.349e-06
```

(c) Confirm your estimates for b0 and b1 from Problem 1

Basd on the data in the summary() command, we note that the estimated b0 and b1 values are 4.6979 and 1.9705, respectively. These values are very close to the b0 and b1 values calculated by hand in Problem 1, which were approximately 4.714 and 1.97, respectively. There is a slight discrepency in both of the values, but this can be attributed to rounding. 1.9705 rounded to the nearest hundreth is 1.97, and if 1.9705 is used in place of 1.97 to calculate b0, we get a value of of 4.6971. This is much closer to the R-estimated value of 4.6979, but this variation can also be attributed to the rounded values for (xbar * ybar), summation(xi*yi), summation(xi^2), and summation(xbar^2). If rounded farther out, such as to the thouandth instead of the hundreth, we could expect the hand calculations and R-calculations to be even closer.

(d) Use the output to report the estimate for variance(epsilon) = theta^2

In the summary() output, we see that the residual standard error is 9.022. Because the residual standard error is the square root of the sample variance, and because the sample variance is also the mean square error, which is also an unbiased estimator for theta^2, then the square of the residual standard error would also be the expected value of epsilon.

Thus, we have v(epsilon) = 9.022^2 = 81.396484

(e) How are the 8 degrees of freedom obtained?

Because there were two unknown variables that were calculated, then the degrees of freedom become n-2. Because the sample size is 10, then the degrees of freedom become:

10 - 8 = 2

Hwk 4-4

Andrew Guo

10/7/2021

4.)

(a) Read the data using the read.csv function

```
crimeLSRL = read.csv("Hwk 4-4.csv")
```

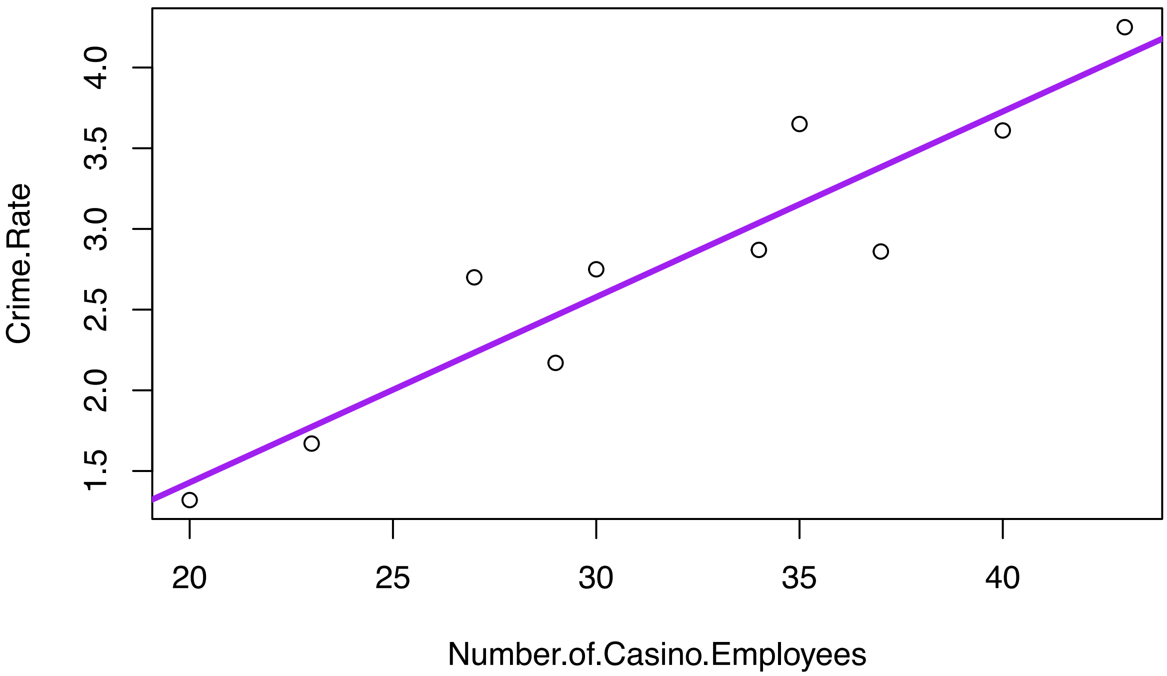
(b) Fit a linear model using the lm function with "Number of Casino Employees" as the predictor and "Crime Rate" as the response. Use the summary function to output the resulting fit.

```
attach(crimeLSRL)
crimeLSRL.fit <- lm(Crime.Rate~Number.of.Casino.Employees, data = crimeLSRL)
summary(crimeLSRL.fit)
```

```
##
## Call:
## lm(formula = Crime.Rate ~ Number.of.Casino.Employees, data = crimeLSRL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5226 -0.1552 -0.1062  0.1763  0.4972
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.86977    0.50903  -1.709    0.126
## Number.of.Casino.Employees  0.11493    0.01564   7.350 7.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3446 on 8 degrees of freedom
## Multiple R-squared:  0.871,  Adjusted R-squared:  0.8549
## F-statistic: 54.03 on 1 and 8 DF,  p-value: 7.992e-05
```

(c) Plot the data along with the regression line

```
plot(Number.of.Casino.Employees, Crime.Rate)
abline(crimeLSRL.fit, lwd = 3, col = 'purple')
```



(d) Use the output to report the estimate for variance(epsilon) = theta^2

In the summary() output, we see that the residual standard error is 0.3446. Because the residual standard error is the square root of the sample variance, and because the sample variance is also the mean square error, which is also an unbiased estimator for theta^2, then the square of the residual standard error would also be the expected value of epsilon.

Thus, we have v(epsilon) = 0.3446^2 = 0.11943836.

(e) Predict the crime rate when there are 25,000 casino employees

Based on our calculations of b1 and b2, our linear regression equation is approximately:

y = 0.115x - 0.867

In Problem 2, we had calculated that the y-intercept (b0) was -0.872, however, this value was reached by using a rounded b1 value of 0.115. In the summary() function, we notice a closer approximation for b1, which is 0.11493. When this number is used in place of 0.115 in the process of solving for b0, we obtain the y-intercept of approximately -0.869. Therefore, the R-calculated value will be used for the intercept.

We are being asked to estimate the crime rate (y) with 25,000 casino employees (x). So we substitute x with its respective value to calculate y. Note that the number of casino employees is denoted as in thousands, so the value being substituted is not 25,000, but rather, 25. Therefore, we calculate y to be:

y = 0.11493(25) - 0.86977 = 2.00348

This means that the expected crime rate when there are 25,000 casino employees would be 2.00348 (per 1,000 population).