STP530 HW3 Solution

2.1

a.

Yes, the conclusion is warranted, because the 95% confidence interval of the true population slope, $(.45, 1.06)$, falls entirely away from zero. That means all plausible values of the true population slope (with 95% confidence) are non-zero. We know that a zero slope indicates the absence of a linear association, so a non-zero slope suggests there is a linear association.

The implied level of significance is $1 - .95 = .05$. Note that for the same underlying model and assumptions, a confidence interval pairs with a corresponding hypothesis testing procedure. Using a 95% confidence interval to conclude whether or not the true population slope equals 0 is equivalent with conducting a t-test of $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$, with the significant level $\alpha = .05$.

b.

It is wise to always pay attention to questionable findings. THEORETICALLY speaking, the intercept means the predicted Y value (dollar sales) when X (number of persons) is 0. So the confidence interval for the intercept gives plausible values of dollar sales when $X = 0$. PRACTICALLY speaking, indeed dollar sales cannot be negative. One may consider replacing the lower bound of the confidence interval with 0.

But the real issue here is extrapolation. A fitted linear regression is only valid within the range of the given X data. For this problem it is reasonable to assume that the given X (million persons) data are all much larger than 0, so making predictions with the fitted regression model outside the reasonable range (e.g., discussing the intercept as the predicted dollar sales when $X = 0$) is a wrong practice.

2.5

b.

First, we use the code below to estimate the regression equation.

```
setwd("~/Documents/ASU/STP530-YiZheng/HW-HaozhenXu/HW2/21spring")

HW2_5.data <-read.table("CH01PR20.txt")

head(HW2_5.data)

colnames(HW2_5.data) <- c("Y", "X")

# Fit the linear regression model on HW2_5.data

my.mod <- lm(Y ~ X, data = HW2_5.data)

summary(my.mod)
```

```
> summary(my.mod)

Call:
lm(formula = Y ~ X, data = HW2_5.data)

Residuals:
     Min       1Q   Median       3Q      Max
-22.7723  -3.7371   0.3334   6.3334  15.4039

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5802     2.8039  -0.207    0.837
X            15.0352     0.4831  31.123   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.914 on 43 degrees of freedom
Multiple R-squared:  0.9575,    Adjusted R-squared:  0.9565
F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

Now to answer the question in part (b): the five steps of this hypothesis test are given below.

- Assumptions:
$$\varepsilon \; i.i.d. \sim N(0, \sigma^2)$$
- Hypotheses:
$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$
- Test-statistic:
$$t^* = \frac{b_1 - 0}{s\{b_1\}} = \frac{b_1}{s/\sqrt{\sum (X_i - \bar{X})^2}} = 31.123$$
which can be found directly from the Coefficients table in the R summary output.

- P-value: P-value is the probability that we get a b_1 this far from the null hypothesis $\beta_1 = 0$ or even more extreme. From the Coefficients table of the R summary output, we can find the P-value $P_{H_0}\{|t^*| > 31.123\} < 2 \times 10^{-16}$.

- Conclusion: Because P-value $< 2 \times 10^{-16} < \alpha = 0.1$, we reject H_0 at a significance level of 0.1 and conclude that the true population slope is not 0, which means there is a linear association between X and Y.

c.

Yes. In part a, the 90% confidence interval for β_1 does not include 0, which implies we would reject $H_0 : \beta_1 = 0$ with a two-sided test. In part b, we did reject $H_0 : \beta_1 = 0$ at the significance level of $\alpha = 0.1$.

When the confidence level of the confidence interval (e.g., $1 - \alpha = 0.9$) and the significance level of the corresponding hypothesis test (e.g., $\alpha = 0.1$) match each other, the conclusion from the confidence interval method is always consistent with the conclusion from the two-sided hypothesis test.

d.

The five steps of this hypothesis test are given below.

- Assumptions:
$$\varepsilon \; i.i.d. \sim N(0, \sigma^2)$$

- Hypotheses:
$$H_0 : \beta_1 \geq 14 \quad H_1 : \beta_1 < 14$$

- Test-statistic:
$$t^* = \frac{b_1 - 14}{s\{b_1\}} = \frac{15.0352 - 14}{0.4831} = 2.1428$$
where the values for b_1 and $s\{b_1\}$ are found from the Coefficients table of the R summary output.

- P-value: Here we have $t^* = 2.1428$. P-value is the probability that we get a b_1 this far from the null hypothesis $\beta_1 \geq 14$ or even more extreme, which is left-tail probability $P_{H_0}\{t^* < 2.1428\} = 0.9811$. This p-value can be found from the following R code:

```
> pt(q=2.1428, df=43)
[1] 0.9810845
```

- Conclusion: Because P-value $= 0.9811 > \alpha = 0.05$, we FAIL TO REJECT $H_0 : \beta_1 \geq 14$ at a significance level of 0.05 and conclude that the true population slope $\beta_1 \geq 14$, which means that the mean required time on call is expected to increase by more than 14 minutes for each additional copier that is served on a service call. The manufacturer's claim is NOT supported.