STP530 HW2 Solution

2.5

a.

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

```
setwd("~/Documents/ASU/STP530-YiZheng/HW-HaozhenXu/HW2/22fall")

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
```

The estimation of the chance in the mean service time when the number of copiers serviced increases by one is given by the slope coefficient b_1. By (2.15), the 90% confidence interval is $b_1 \pm t(1-\alpha/2; n-p) \cdot s\{b_1\} = 15.0352 \pm 1.6811 \times 0.4831$, which is $[14.2231, 15.8473]$. See more detail below.

$b_1 = 15.0352$ is the estimate of the slope in the regression model, which can be found in the R output.

$s\{b_1\} = 0.4831$ can be either found in the R output under the column of Std.Error, or derived by $s\sqrt{\frac{1}{\sum(X_i - \bar{X})^2}}$.

With $n = 45$, $p = 2$, $\alpha = 1 - 90\% = 0.1$, the 90% critical t-value $t(1-\alpha/2 = .95; df = n - p = 45 - 2 = 43) = 1.6811$, which can be found by the following R code:

```
> qt(p=.95, df=43)
[1] 1.681071
```

There is a function that provides the confidence intervals of the regression parameters directly:

```
> confint(my.mod, level=.90)
                 5 %      95 %
(Intercept) -5.29378  4.133467
X           14.22314 15.847352
```

Interpretation: With a 90% confidence, we estimate that the mean service time increases by somewhere between 14.2231 minutes and 15.8473 minutes when the number of copiers serviced increases by one.

e.

The regression intercept b_0 is -0.5802 based on the R output. However, like in problem 2.1 (b), when $X = 0$ is outside of the range of the given data, it is called extrapolation, and the fitted regression equation does not necessarily hold. So there is little information given by b_0.

2.14.

a.

By (2.33) and (2.30), the 90% confidence interval for the mean service time on calls in which 6 copiers are serviced, $E\{Y_h|X_h = 6\}$ is:

$$\hat{Y}_h \pm t(1-\alpha/2; n-p) \cdot s\{\hat{Y}_h\} = (b_0 + b_1 X_h) \pm t(1-\alpha/2; n-p) \cdot s\sqrt{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}},$$

where $X_h = 6$, $n = 45$, $p = 2$, $\alpha = 1 - 90\% = .1$.

From the R output we obtained in Question 2.5, we have $b_0 = -0.5802$, $b_1 = 15.0352$, and $s = 8.914$. The critical t-value $t(1-\alpha/2; n-p)$ is obtained as `t.crit` in the R code below. Thus the 90% confidence interval for $E\{Y_h|X_h = 6\}$ is [87.28, 91.98].

```
> attach(HW2_5.data)
> n <- nrow(HW2_5.data)
> p <- 2
> X.h <- 6
> s <- 8.914    # You can also get a more accurate s by summary(my.mod)$sigma
>
> Y.hat <- -0.5802 + 15.0352 * X.h
> t.crit <- qt(p = .95, df = n - p)
> se.CI.Y.hat <- s * sqrt(1/n + (X.h - mean(X)) ^ 2 / sum((X - mean(X)) ^ 2))
>
> # Lower limit
> Y.hat - t.crit * se.CI.Y.hat
[1] 87.28341
>
> # Upper limit
> Y.hat + t.crit * se.CI.Y.hat
[1] 91.97859
```

We can also obtain the confidence interval using the following R code, which gives a more precise result (less rounding error):

```
> predict(my.mod, newdata=data.frame(X=6), interval="confidence", level=.90)
       fit      lwr     upr
1 89.63133 87.28387 91.9788
```

Interpretation: With confidence coefficient 0.90, we estimate that the mean service time on calls in which 6 copier are serviced is some value between 87.2838 minutes and 91.9788 minutes.

b.

By (2.36) and (2.38a), the 90% prediction interval for the service time on the next call in which 6 copiers are serviced is:

$$\hat{Y}_h \pm t(1-\alpha/2; n-p)s\{pred\} = (b_0+b_1X_h) \pm t(1-\alpha/2; n-p) \cdot s\sqrt{1+\frac{1}{n}+\frac{(X_h-\bar{X})^2}{\sum_i (X_i-\bar{X})^2}}.$$

where $X_h = 6$, $n = 45$, $p = 2$, $\alpha = 1 - 90\% = .1$.

Based on the above formula, we can obtain the prediction interval for $Y_h|X_h = 6$ as [74.46, 104.80]. The R codes below are additional codes to what was run in part (a) in order to answer this question.

```
> se.PI.Y.hat <- s * sqrt(1 + 1/n + (X.h - mean(X)) ^ 2 / sum((X - mean(X)) ^ 2))
>
> # Lower limit
> Y.hat - t.crit * se.PI.Y.hat
[1] 74.46316
>
> # Upper limit
> Y.hat + t.crit * se.PI.Y.hat
[1] 104.7988
```

We can also obtain the prediction interval using the following R code:

```
> predict(my.mod, newdata=data.frame(X=6), interval="prediction", level=.90)
       fit      lwr      upr
1 89.63133 74.46433 104.7983
```

Interpretation: With a 90% confidence, we predict the service time on the next call in which six copiers are serviced is somewhere between 74.46 minutes and 104.80 minutes.

Yes the prediction interval is wider than the corresponding confidence interval, and it should be. The reason is, the prediction interval for a new observation counts for not only the sampling variability of the point estimator $\hat{Y}_h$, which is the variance of the sampling distribution of $\hat{Y}_h$, but also the error variance of the new observation itself, which is σ^2, the variance of the error term in the assumed linear model. To make it clear, we just need to recall the regression function:

$$Y_h = \beta_0 + \beta_1 X_h + \varepsilon = E\{Y_h\} + \varepsilon,$$

and then, because $E\{Y_h\}$ and ε are independent, we have

$$var\{Y_h\} = var\{E\{Y_h\}\} + \sigma^2.$$

The point estimator of $E\{Y_h\}$ and the prediction of the new observation Y_h are both $\hat{Y}_h = b_0 + b_1 X_h$. However, when it comes to confidence/prediction intervals, the prediction variance of Y_h is the summation of the sampling variance of $E\{Y_h\}$ and the error variance σ^2 as shown in the equation above. A larger variance means a larger standard error, which leads to a wider interval.