STP530 HW4 Solution

2.25

a.

See the R codes below to calculate SSR, SSE, and SSTO in the ANOVA table. Follow the example in textbook Table 2.2 to construct the ANOVA table.

```
# Set the working directory to where you saved the data file

setwd("~/Documents/ASU/STP530-YiZheng/HW-HaozhenXu/HW4")

HW2_25.data <- read.table("CH01PR21.txt")
colnames(HW2_25.data) <- c("Y", "X")
head(HW2_25.data)

mod1 <- lm(Y ~ X, data=HW2_25.data)
summary(mod1)

# Calculate SSTO

SSR <- sum((predict(mod1) - mean(HW2_25.data$Y)) ^ 2)
SSE <- sum((HW2_25.data$Y - predict(mod1)) ^ 2)
SSTO <- sum((HW2_25.data$Y - mean(HW2_25.data$Y)) ^ 2)

# You can also use the anova() function in R to see a part of the ANOVA table

anova(mod1)
```

See R outputs below.

```
> SSR
[1] 160
> SSE
[1] 17.6
> SSTO
[1] 177.6
```

```
> anova(mod1)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1  160.0   160.0  72.727 2.749e-05 ***
Residuals  8   17.6     2.2
```

Now we can construct the ANOVA table as below.

Source of Variation	***SS***	***df***	***MS***
Regression	160.0	1	160.0
Error	17.6	8	2.2
Total	177.6	9	19.7

From the table above, we can see that the sums of square of regression and error and their corresponding degrees of freedom are additive.

b.

Five steps of F test:

- 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:
$$F^* = \frac{MSR}{MSE} = \frac{160}{2.2} = 72.727$$
- P-value: Here the P-value is $P_{H_0}\{F^* > 72.727\} = 2.75 \times 10^{-5}$. Use R code `1 - pf(q=72.727, df1=1, df2=8)`
- Conclusion: Because the P-value $= 2.75 \times 10^{-5} < \alpha = 0.05$, we reject H_0 at a significance level of 0.05. The population slope of this linear model is indeed non-zero.

c.

Using command $summary()$ in R to view results of the fitted model, we have $t^* = 8.528$, and we can find $t^{*2} = 8.529^2 = 72.7 = F^*$ in (b). Also notice the p-values of these two tests are identical.

```
> summary(mod1)

Call:
lm(formula = Y ~ X, data = hwdata1)

Residuals:
   Min     1Q Median     3Q    Max
  -2.2   -1.2    0.3    0.8    1.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.2000     0.6633  15.377 3.18e-07 ***
X             4.0000     0.4690   8.528 2.75e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.483 on 8 degrees of freedom
Multiple R-squared:  0.9009,    Adjusted R-squared:  0.8885
F-statistic: 72.73 on 1 and 8 DF,  p-value: 2.749e-05
```

d.

We can calculate R^2 by using the values we entered in the ANOVA table:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} = 1 - \frac{17.6}{177.6} = 0.9009$$

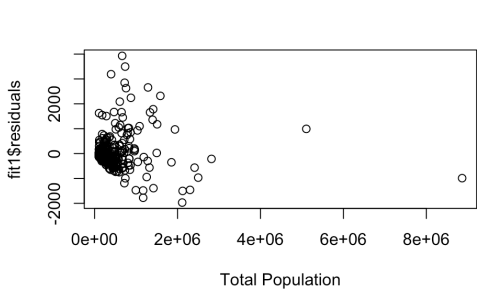
Also from the result in R, we have that $R^2 = 0.9009$, and thus $r = \sqrt{R^2} = 0.9492$. By the definition of R^2, 90.09% of the variation in the response Y is explained by this fitted model.

3.25

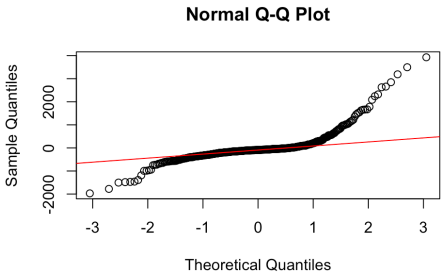
We can use the code below to fit models and get residual plots and Q-Q plots respectively.

```
HW3_25.data <- read.table("APPENC02.txt")
head(HW3_25.data)

# change column names

colnames(HW3_25.data)[8] <- "Y"
colnames(HW3_25.data)[5] <- "X1"
colnames(HW3_25.data)[9] <- "X2"
colnames(HW3_25.data)[16] <- "X3"

#fit three models respectively

fit1<-lm(Y ~ X1, data = HW3_25.data)
fit2<-lm(Y ~ X2, data = HW3_25.data)
fit3<-lm(Y ~ X3, data = HW3_25.data)

# obtain residuals and residual plots and Q-Q plot respectively

plot(HW3_25.data$X1, fit1$residuals, xlab="Total Population")
qqnorm(fit1$residuals)
qqline(fit1$residuals, col='red')

plot(HW3_25.data$X2, fit2$residuals, xlab="Number of Hospital Beds")
qqnorm(fit2$residuals)
qqline(fit2$residuals, col='red')

plot(HW3_25.data$X3, fit3$residuals, xlab="Total Personal Income")
qqnorm(fit3$residuals)
qqline(fit3$residuals, col='red')
```

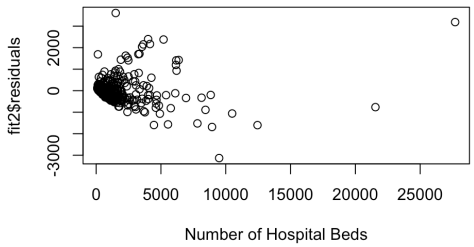
See the figures on the next page.



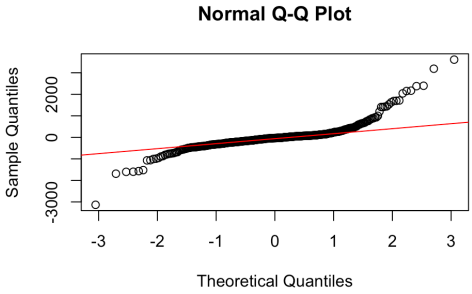
(a) Total Population Residual Plot



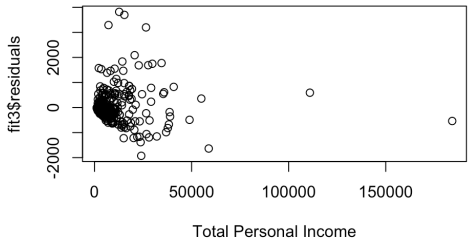
(b) Total Population Q-Q plot



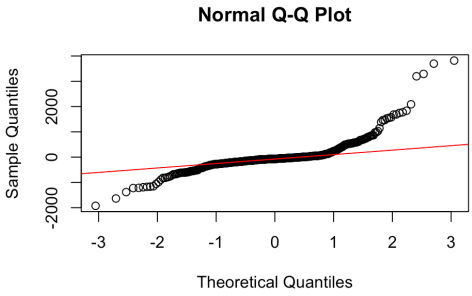
(c) Number of Hospital Beds Residual Plot



(d) Number of Hospital Beds Q-Q plot



(e) Total Personal Income Residual Plot



(f) Total Personal Income Q-Q Plot

Figure 1: Residual Plots and Q-Q Plots

From the figures above, none of the three simple linear regression models seems a very good fit. For all three models, the residuals tend to distribute with a wider range when the value of X increases (i.e., the funnel shape), and the Q-Q plots are not very close to the straight reference lines on both ends.

None of the three models looks obviously more appropriate than the others.