STP530 HW6 Solution

7.5

a.

We can use the R codes below to obtain the required ANOVA table.

```
setwd("~/Documents/ASU/STP530-YiZheng/HW-HaozhenXu/HW6")

# Import dataset

HW6.data <- read.table("CH06PR15.txt")
colnames(HW6.data) <- c("Y", "X1", "X2", "X3")

# Fit the linear regression model. The order of the predictors in the
# lm() function must follow the order of entrance specified by the problem
# for anova() to return the correct extra sum of squares.

mod <- lm(Y ~ X2 + X1 + X3, data=HW6.data)

# Generate the anova table

anova(mod)
```

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

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X2         1 4860.3  4860.3 48.0439 1.822e-08 ***
X1         1 3896.0  3896.0 38.5126 2.008e-07 ***
X3         1  364.2   364.2  3.5997   0.06468 .
Residuals 42 4248.8   101.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We take the respective values from the output of `anova()` to construct the following ANOVA table. Note that the last row "Total" is not directly given by `anova()`. We can sum up the SS and df columns to obtain the total SS and total df.

Source of Variation	***SS***	***df***	***MS***
X_2	4860.3	1	4860.3
$X_1 \mid X_2$	3896.0	1	3896.0
$X_3 \mid X_1, X_2$	364.2	1	364.2
Error	4248.8	42	101.2
Total	13369.3	45	

b.

We use the general linear test approach to test whether X_3 can be dropped from the model.

- Step 1. Assumptions:
$$\varepsilon \; i.i.d. \sim N(0, \sigma^2)$$
- Step 2. Hypotheses:
$$\text{The full model: } E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$
$$\text{The reduced model: } E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$
$$H_0 : \beta_3 = 0 \quad H_1 : \beta_3 \neq 0$$
- Step 3. Test-statistic: (All relevant quantities are directly available in the ANOVA table above.)
$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F} = \frac{SSR(X_3|X_1, X_2)}{df_R - df_F} \div \frac{SSE(X_1, X_2, X_3)}{df_F}$$
$$= \frac{MSR(X_3|X_1, X_2)}{MSE(X_1, X_2, X_3)} = \frac{364.2}{101.2} = 3.5997$$
- Step 4. P-value: Here the P-value is the right-tail probability on the reference distribution $F(df1 = 1, df2 = 42)$, which is $P_{H_0}\{F^* > 3.5997\} = 0.065$. This value is also given in the R output of `anova()`.
- Step 5. Conclusion: Note that the level of significance required by this problem is $\alpha = .025$. Because the P-value $> \alpha = 0.025$, we can not reject H_0 at the significance level of 0.025, which means the full model does not fit the data significantly better than the reduced model, so we should go with the reduced model (i.e., we CAN drop X_3 from the model). Putting this in context, with the patient's age and severity of illness already in the model, anxiety level does not further enhance the prediction of the patient's satisfaction level.

7.6

This problem asks whether both X2 and X3 can be dropped from the model given that X1 is retained in the model. This question can be answered by carrying out a general linear test approach to compare the full model, where Y is predicted by X1, X2, and X3, vs. the reduced model, where Y is predicted by X1 only. With the following R codes, we specify the full and reduced models and use `anova` to obtain the SSE and df for each model.

```
mod.F <- lm (Y ~ X1 + X2 + X3, data=HW6.data)
anova(mod.F)
mod.R <- lm (Y ~ X1, data=HW6.data)
anova(mod.R)
```

See R outputs below:

```
> mod.F <- lm (Y ~ X1 + X2 + X3, data=HW5.data)
> anova(mod.F)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 8275.4  8275.4 81.8026 2.059e-11 ***
X2         1  480.9   480.9  4.7539   0.03489 *
X3         1  364.2   364.2  3.5997   0.06468 .
Residuals 42 4248.8   101.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> mod.R <- lm (Y ~ X1, data=HW5.data)
> anova(mod.R)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 8275.4  8275.4  71.481 9.058e-11 ***
Residuals 44 5093.9   115.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The five steps of the F test:

- Step 1. Assumptions:

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

- Step 2. Hypotheses:

$$\text{The full model: } E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$\text{The reduced model: } E\{Y\} = \beta_0 + \beta_1 X_1$$

$$H_0 : \beta_2 = \beta_3 = 0 \quad H_1 : \text{not both } \beta_2 \text{ and } \beta_3 \text{ equal } 0$$

- Step 3. Test-statistic: (Find the quantities from the R outputs above.)

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$
$$= \frac{5093.9 - 4248.8}{44 - 42} \div \frac{4248.8}{42} = 4.177$$

- Step 4. P-value: Here the P-value is the right-tail probability on the reference distribution $F(df1 = 2, df2 = 42)$, which is $P_{H_0}\{F^* > 4.1768\} = 0.022$. This value can be obtained by R codes: `1 - pf(q=4.177, df1=2, df2=42)`.

- Step 5. Conclusion: Because the P-value $< \alpha = 0.025$, we can reject H_0 at the significance level of 0.025, which means the full model fits the data significantly better than the reduced model, so we CANNOT drop both X_2 and X_3 from the full model. Putting this in context, the combination of the severity of illness and anxiety level does further explain the patients' satisfaction level after what the patient's age accounts for. (Of course, from the previous problem, we know that anxiety level is redundant once the severity of illness is included.)

You can also use the following simple R codes to compare two nested models using the general linear test approach.

```
> mod.F <- lm (Y ~ X1 + X2 + X3, data=HW5.data)
> mod.R <- lm (Y ~ X1, data=HW5.data)
> anova(mod.F, mod.R)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X3
Model 2: Y ~ X1
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     42 4248.8
2     44 5093.9 -2   -845.07 4.1768 0.02216 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

7.9

The problem asks whether $\beta_1 = -1$ and $\beta_2 = 0$. We can answer this question by using the general linear test approach, where we set up the **full model** to be:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

And we set up the **reduced model** to include those two parameter constraints:

$$E\{Y\} = \beta_0 + (-1)X_1 + (0)X_2 + \beta_3 X_3$$

And the above reduced model is equivalent with:

$$E\{Y\} + X_1 = \beta_0 + \beta_3 X_3$$

Using the R codes below, we can obtain the SSE and df of each model to compute the test statistic.

```
> mod.F <- lm (Y ~ X1 + X2 + X3, data=HW5.data)
> anova(mod.F)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 8275.4  8275.4 81.8026 2.059e-11 ***
X2         1  480.9   480.9  4.7539   0.03489 *
X3         1  364.2   364.2  3.5997   0.06468 .
Residuals 42 4248.8   101.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> mod.R <- lm (Y + X1 ~ X3, data=HW5.data)
> anova(mod.R)
Analysis of Variance Table

Response: Y + X1
          Df Sum Sq Mean Sq F value    Pr(>F)
X3         1 1636.3 1636.26   16.26 0.0002162 ***
Residuals 44 4427.7  100.63
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The five steps of F test,

- Step 1. Assumptions:

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

- Step 2. Hypotheses:

$$\text{The full model: } E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$\text{The reduced model: } E\{Y\} + X_1 = \beta_0 + \beta_3 X_3$$

$$H_0 : \beta_1 = -1, \beta_2 = 0 \quad H_1 : \text{not both equalities in } H_0 \text{ hold}$$

- Step 3. Test-statistic:

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F} = \frac{4427.7 - 4248.8}{44 - 42} \div \frac{4248.8}{42} = 0.884$$

- Step 4. P-value: Here the P-value is the right-tail probability on the reference distribution $F(df1 = 2, df2 = 42)$, which is $P_{H_0}\{F^* > 0.884\} = 0.4207$. This value can be obtained by R codes: `1 - pf(q=0.884, df1=2, df2=42)`.

- Step 5. Conclusion: Because the P-value $> \alpha = 0.025$, we can not reject H_0 at the significance level of 0.025, which means the full model doesn't fit the data significantly better than the reduced model, so we can use the reduced model, which means it's possible that $\beta_1 = -1$ and $\beta_2 = 0$.

See below for additional notes.

```
> # Attempting to use the following codes to conduct the general linear test approach results in an error.
>
> mod.F <- lm (Y ~ X1 + X2 + X3, data=HW5.data)
> mod.R <- lm (Y + X1 ~ X3, data=HW5.data)
> anova(mod.F, mod.R)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 8275.4  8275.4 81.8026 2.059e-11 ***
X2         1  480.9   480.9  4.7539   0.03489 *
X3         1  364.2   364.2  3.5997   0.06468 .
Residuals 42 4248.8   101.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Warning message:
In anova.lmlist(object, ...) :
  models with response ‘"Y + X1"’ removed because response differs from model 1
>
> # We can get around the error by manually compute a new response variable that reflects the reduced model.
>
> new.data <- HW5.data
> new.data$Y <- HW5.data$Y + HW5.data$X1
> mod.F <- lm (Y ~ X1 + X2 + X3, data=HW5.data)
> mod.R.new <- lm (Y ~ X3, data=new.data)
> anova(mod.F, mod.R.new)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X3
Model 2: Y ~ X3
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     42 4248.8
2     44 4427.7 -2   -178.81 0.8838 0.4208
```

7.14

a.

To obtain the coefficients of partial determination, we need to know the SSEs and extra sums of squares of a variety of models. See below for the R codes.

```
> m1 <- lm (Y ~ X1, data=HW5.data)
> anova(m1)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 8275.4  8275.4  71.481 9.058e-11 ***
Residuals 44 5093.9   115.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> m2 <- lm (Y ~ X2, data=HW5.data)
> anova(m2)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
X2         1 4860.3  4860.3  25.132 9.23e-06 ***
Residuals 44 8509.0   193.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> m21 <- lm (Y ~ X2 + X1, data=HW5.data)
> anova(m21)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X2         1 4860.3  4860.3  45.305 3.161e-08 ***
X1         1 3896.0  3896.0  36.317 3.348e-07 ***
Residuals 43 4613.0   107.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> m23 <- lm (Y ~ X2 + X3, data=HW5.data)
> anova(m23)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X2         1 4860.3  4860.3 29.4089 2.507e-06 ***
X3         1 1402.7  1402.7  8.4873  0.005653 **
Residuals 43 7106.4   165.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> m231 <- lm (Y ~ X2 + X3 + X1, data=HW5.data)
> anova(m231)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X2         1 4860.3  4860.3  48.044 1.822e-08 ***
X3         1 1402.7  1402.7  13.865 0.0005788 ***
X1         1 2857.6  2857.6  28.247 3.810e-06 ***
Residuals 42 4248.8   101.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that in each `anova()` output above, the SS terms give the extra sums of squares where the predictors are entered by the order listed. The last row gives the SSE of the respective model. Summing up all SS terms in each table will give us the same SSTO value.

We now take relevant values from the outputs to compute the three coefficients of partial determination:

$$R^2_{Y1} = 1 - \frac{SSE(X_1)}{SSTO} = 1 - \frac{5093.9}{13369.3} = 0.6190$$
$$R^2_{Y1|2} = \frac{SSR(X_1|X_2)}{SSE(X_2)} = \frac{3896.0}{8509.0} = 0.4579$$
$$R^2_{Y1|23} = \frac{SSR(X_1|X_2, X_3)}{SSE(X_2, X_3)} = \frac{2857.6}{7106.4} = 0.4021$$

Comparing these three partial R^2's, we can find when adjusted for X_2 or both X_2 and X_3, the degree of marginal linear association between Y and X_1 reduces, but not much.

b.

Following the same process, we run the models given on the next page to obtain the SSEs and extra sums of squares of the relevant models and then compute the partial R^2s.

$$R^2_{Y2} = 1 - \frac{SSE(X_2)}{SSTO} = 1 - \frac{8509.0}{13369.3} = 0.3635$$
$$R^2_{Y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{480.9}{5093.9} = 0.0944$$
$$R^2_{Y2|13} = \frac{SSR(X_2|X_1, X_3)}{SSE(X_1, X_3)} = \frac{81.7}{4330.5} = 0.0189$$

Contrary to part (a), when adjusted for X_1 or both X_1 and X_3, the degree of marginal linear association between Y and X_2 reduces significantly, from a moderate association to a rather weak one.

The combined pattern seen in the two parts suggests that X_1 has the strongest tie to Y, whereas X_2's seeming marginal association with Y may be superficial or may be mediated through X_1. This pattern provides evidence supporting a plausible causal relationship between X_1 (patient's age) and Y (patient satisfaction), or that patient's age is the biggest reason explaining his/her dissatisfaction of the hospital. On the other hand, the seeming association between severity of illness (X_2) and patient's satisfaction level (Y) mostly reflects the association between patient's age (X_1) and patient's satisfaction level (Y), leaving only a slim amount of unique association.

```
> m2 <- lm (Y ~ X2, data=HW5.data)
> anova(m2)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X2         1 4860.3  4860.3  25.132 9.23e-06 ***
Residuals 44 8509.0   193.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> m1 <- lm (Y ~ X1, data=HW5.data)
> anova(m1)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 8275.4  8275.4  71.481 9.058e-11 ***
Residuals 44 5093.9   115.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> m12 <- lm (Y ~ X1 + X2, data=HW5.data)
> anova(m12)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 8275.4  8275.4 77.1389 3.802e-11 ***
X2         1  480.9   480.9  4.4828   0.04006 *
Residuals 43 4613.0   107.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> m13 <- lm (Y ~ X1 + X3, data=HW5.data)
> anova(m13)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 8275.4  8275.4 82.1711 1.557e-11 ***
X3         1  763.4   763.4  7.5804   0.00861 **
Residuals 43 4330.5   100.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> m132 <- lm (Y ~ X1 + X3 + X2, data=HW5.data)
> anova(m132)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 8275.4  8275.4 81.8026 2.059e-11 ***
X3         1  763.4   763.4  7.5464  0.008819 **
X2         1   81.7    81.7  0.8072  0.374070
Residuals 42 4248.8   101.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```