Homework #6

Dustin Leatherman

October 28, 2019

7.1

State the number of the degrees of freedom associated with the following

1

$SSR(X_1|X_2)$

df: 1

2

$SSR(X_2|X_1, X_3)$

df: 1

3

$SSR(X_1, X_2|X_3, X_4)$

df: 2

4

$SSR(X_1, X_2, X_3|X_4, X_5)$

df: 3

7.2

Explain in what sense the regression some of squares $SSR(X_1)$ is an extra sum of squares

The Extra Sum of Squares is defined as the marginal reduction in SSE when one or several predictors are added to the regression model, *given* other predictors are already in the model. $SSR(X_1)$ is the effect of X_1 in the model given the other predictors are constant.

7.28

a

Define each of the following extra sums of squares

$SSR(X_5|X_1)$

$SSR(X_5|X_1) = SSR(X_1, X_5) - SSR(X_1)$

$SSR(X_3, X_4|X_1)$

$SSR(X_3, X_4|X_1) = SSR(X_1, X_3, X_4) - SSR(X_1)$

$SSR(X_4, |X_1, X_2, X_3)$

$SSR(X_4, |X_1, X_2, X_3) = SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2, X_3)$

b

For a multiple regression models with 5 X variables, what is the relevant sum of squares for testing if $\beta_5 = 0$? whether or not $\beta_2 = \beta_4 = 0$?

$\beta_5 = 0 \rightarrow SSR(X_5|X_1, X_2, X_3, X_4) = SSR(X_1, X_2, X_3, X_4, X_5) - SSR(X_1, X_2, X_3, X_4)$

$\beta_2 = \beta_4 = 0 \rightarrow SSR(X_2, X_4|X_1, X_3, X_5) = SSR(X_1, X_2, X_3, X_4, X_5) - SSR(X_1, X_3, X_5)$

7.29a

Show that $SSR(X_1, X_2, X_3, X_4) = SSR(X_1) + SSR(X_2, X_3|X_1) + SSR(X_4|X_1, X_2, X_3)$

1. Define the Extra Sums of Squares

   $SSR(X_2, X_3|X_1) = SSR(X_1, X_2, X_3) - SSR(X_1) SSR(X_4|X_1, X_2, X_3) = SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2, X_3)$

2. Substitute and Reduce terms

$SSR(X_1, X_2, X_3, X_4) = SSR(X_1) + [SSR(X_1, X_2, X_3) - SSR(X_1)] + [SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2, X_3)] = SSR(X_1, X$

8.8

Refer to the commercial properties problems. The vacancy rate predictor X_3 does not appear to be needed when property age X_1, operating expenses and taxes X_2, and total square footage X_4 are included in the model as predictors of rental rates Y.

a

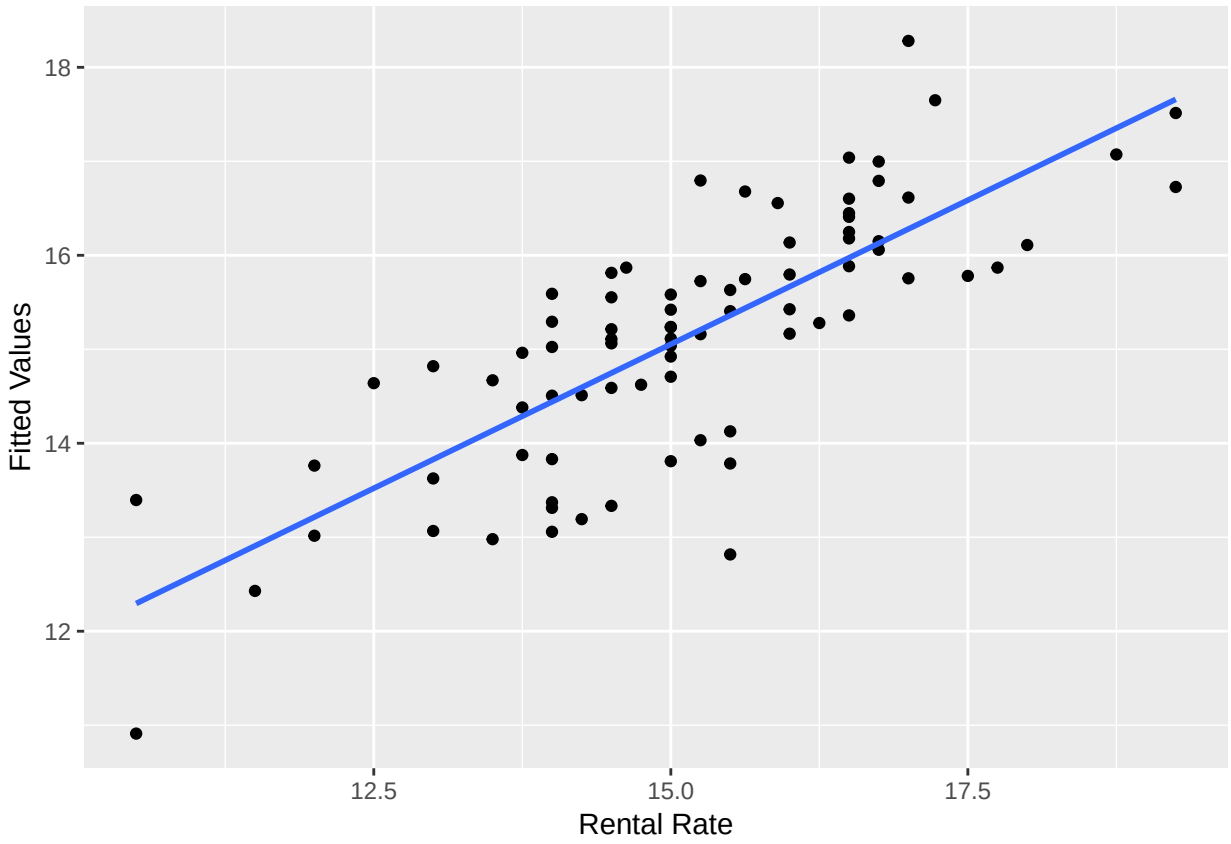
The age of property (X_1) appears to exhibit some curvature when plotted against rental rates (Y). Fit a polynomial regression model with centered property age (X_1), the square of centered property age (X_1^2), operating expenses and taxes (X_2), and total square footage (X_4). Plot the Y observations against fitted values. Does the response function provide a good fit?

```
properties <- read.csv("~/snap/firefox/common/Downloads/commercialproperties.csv")

# calculate centered age
properties$age.center <- scale(properties$age, scale = FALSE)

properties.model <- lm(rentalrate ~ poly(age.center, 2) + operatingexpense + squarefootage, data = prope

qplot(rentalrate, .fitted, data = properties.model, xlab = "Rental Rate", ylab = "Fitted Values") + geo
```



The response function provides an okay fit. There is an uneven amount of fitted values above and below the line indicating that this model leaves something to be desired. This is confirmed by the adjusted R^2 value being 0.5927. Further models should be considered.

c

Test whether or not the square of centered age can be dropped from the model. use $\alpha = 0.05$. State the alternatives, decision rule, and conclusion. What is the p-value of the test?

$H_0 : X_1^2 = 0$ $H_A : X_1^2 \neq 0$

```
properties.model.linear <- lm(rentalrate ~ poly(age.center, 1) + operatingexpense + squarefootage, data

anova(properties.model.linear, properties.model) %>% kable
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
77	98.65034	NA	NA	NA	NA
76	91.53496	1	7.115385	5.90779	0.0174321

There is moderate evidence that the squared of the center age is significant in this model (Sum of Squares F-Test. p-value = 0.01743). Thus, this predictor should not be dropped.

1

Carry out all hypothesis tests at 5% significance level. Consider brand preference data. Let Moisture be X_1 and Sweetness be X_2

a

Obtain and Interpret $SSR(X_1|X_2)$ and $SSR(X_2|X_1)$

```
brands <- read.csv("~/snap/firefox/common/Downloads/brandperference.csv")

brands.model <- lm(liking ~ sweetness + moisture, data = brands)
anova(brands.model) %>% kable
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sweetness	1	306.25	306.250000	42.21898	2.01e-05
moisture	1	1566.45	1566.450000	215.94751	0.00e+00
Residuals	13	94.30	7.253846	NA	NA

$SSR(X_1|X_2) = SSR(X_1, X_2) - SSR(X_2) = 1566.45 + 306.25 - 306.25 = 1566.45$

The marginal effect of the extra sum of squares for moisture after accounting for sweetness is 1566.45.

```
brands.model2 <- lm(liking ~ moisture + sweetness, data = brands)
anova(brands.model2) %>% kable
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
moisture	1	1566.45	1566.450000	215.94751	0.00e+00
sweetness	1	306.25	306.250000	42.21898	2.01e-05
Residuals	13	94.30	7.253846	NA	NA

$SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1) = 1566.45 + 306.25 - 1566.25 = 306.25$

The marginal effect of the extra sum of squares for sweetness after accounting for moisture is 306.25.

b

Obtain $SSR(X_1), SSR(X_2|X_1)$ and verify $SSR(X_1, X_2) = SSR(X_1) + SSR(X_2|X_1)$

$SSR(X_1) = 1566.45$

$SSR(X_2|X_1) = 306.25$

$SSR(X_1, X_2) = SSR(X_1) + SSR(X_2|X_1) = 1872.7$

c

Obtain and Interpret $R^2_{Y|12}$ and $R^2_{Y|21}$

$R^2_{Y|12} = \frac{SSR(X_1|X_2)}{SSE(X_2)} = \frac{1566.45}{1566.45+94.3} = 0.9432$

The proportion of variation explained by moisture given sweetness is already in the model is 94.32%.

$R^2_{Y|21} = \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{306.25}{306.25+94.3} = 0.7646$

The proportion of variation explained by sweetness given moisture is already in the model is 76.46%.

d

Compute the correlation matrix between X_1 and X_2. Comment on how strongly correlated the two predictors are. Should we be concerned with multicollinearity in the model?

```
cor(brands %>% select(-liking)) %>% kable
```

	moisture	sweetness
moisture	1	0
sweetness	0	1

There is no correlation between moisture and sweetness so there is no concern for multicollinearity in this model.

2

Consider commercial properties data. Let age: X_1, operatingexpense: X_2, vacancy: X_3, squarefootage: X_4

```
properties <- read.csv("~/snap/firefox/common/Downloads/commercialproperties.csv")
properties.model <- lm(rentalrate ~ ., data = properties)
```

a

Obtain and Interpret $SSR(X_1), SSR(X_2|X_1), SSR(X_3|X_2, X_1), SSR(X_4|X_1, X_2, X_3)$

$SSR(X_1) = 14.819$

The marginal effect of the extra sum of squares for age in the model is 14.819.

$SSR(X_2|X_1) = 72.802$

The marginal effect of the extra sum of squares for operating expenses after accounting for age is 72.802.

$SSR(X_3|X_1, X_2) = 8.381$

The marginal effect of the extra sum of squares for vacancy after accounting for age and operating expenses is 8.381.

$SSR(X_4|X_1, X_2, X_3) = 42.325$

The marginal effect of the extra sum of squares for square footage after accounting for age, operating expenses, and vacancy is 42.325.

b

Verify that the above extra sum of squares in (a) sum to SSR(X_1, X_2, X_3, X_4)

$SSR(X_1, X_2, X_3, X_4) = SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2) + SSR(X_4|X_1, X_2, X_3) = 14.819 + 72.802 + 8.381 + 42.325 = 138.327$

c

Obtain and Interpret $R^2_{Y1|234}$, $R^2_{Y2|134}$, $R^2_{Y3|214}$, and $R^2_{Y4|231}$

```
# Use Anova() to retrieve Type II SS to avoid re-running anova() with different orders of predictors.
car::Anova(properties.model) %>% kable
```

	Sum Sq	Df	F value	Pr(>F)
age	57.2427624	1	44.2881364	0.0000000
operatingexpense	25.7589552	1	19.9294387	0.0000275

	Sum Sq	Df	F value	Pr(>F)
vacancy	0.4197463	1	0.3247534	0.5704457
squarefootage	42.3249580	1	32.7463846	0.0000002
Residuals	98.2305939	76	NA	NA

$R^2_{Y1|234} = \frac{SSR(X_1|X_2,X_3,X_4)}{SSE(X_2,X_3,X_4)} = \frac{57.243}{57.243+98.231} = 0.3682$

The proportion of variation explained by age given operating expense, vacancy, and square footage are already in the model is 36.82%.

$R^2_{Y2|134} = \frac{SSR(X_2|X_1,X_3,X_4)}{SSE(X_1,X_3,X_4)} = \frac{25.759}{25.759+98.231} = 0.2078$

The proportion of variation explained by operating expenses given age, vacancy, and square footage are already in the model is 20.78%.

$R^2_{Y3|214} = \frac{SSR(X_3|X_2,X_1,X_4)}{SSE(X_2,X_1,X_4)} = \frac{0.420}{0.420+98.231} = 0.0043$

The proportion of variation explained by vacancy given operating expense, age, and square footage are already in the model is 0.43%.

$R^2_{Y4|231} = \frac{SSR(X_4|X_2,X_3,X_1)}{SSE(X_2,X_3,X_1)} = \frac{42.325}{42.325+98.231} = 0.3011$

The proportion of variation explained by square footage given operating expense, vacancy, and age are already in the model is 30.11%.

d

> Compute the correlation matrix for X_1, X_2, X_3, X_4. Comment on how strongly correlated the two predictors are. Should we be concerned with multicollinearity in the model?

```
cor(properties %>% select(-rentalrate)) %>% kable
```

	age	operatingexpense	vacancy	squarefootage
age	1.0000000	0.3888264	-0.2526635	0.2885835
operatingexpense	0.3888264	1.0000000	-0.3797617	0.4406971
vacancy	-0.2526635	-0.3797617	1.0000000	0.0806107
squarefootage	0.2885835	0.4406971	0.0806107	1.0000000

The max absoluate correlation between age, vacancy, square footage, and operating expense is 0.4407 (operating expense and square footage). This is a very weak correlation and thus should not be a concern with regards to the model.