STP 530: Applied Regression Analysis

Name : **Sai Swaroop Reddy Vennapusa**

Homework **9**

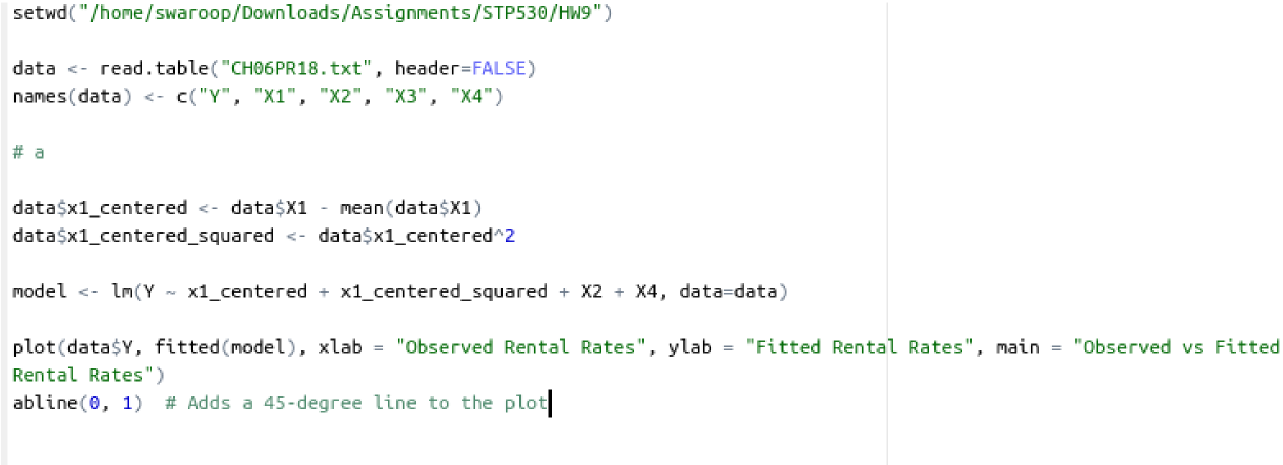
Instructor : **Yi Zheng**

Due Date : 7th Nov 2023, 10:30AM

Question 8.8 a. The age of the property (X1) appears to exhibit some curvature when plotted against the rental rates (Y). Fit a polynomial regression model with centered property age (x1), the square of centered property age (x1^2), operating expenses and taxes (X2), and total square footage (X4). Plot the Y observations against the fitted values. Does the response function provide a good fit?

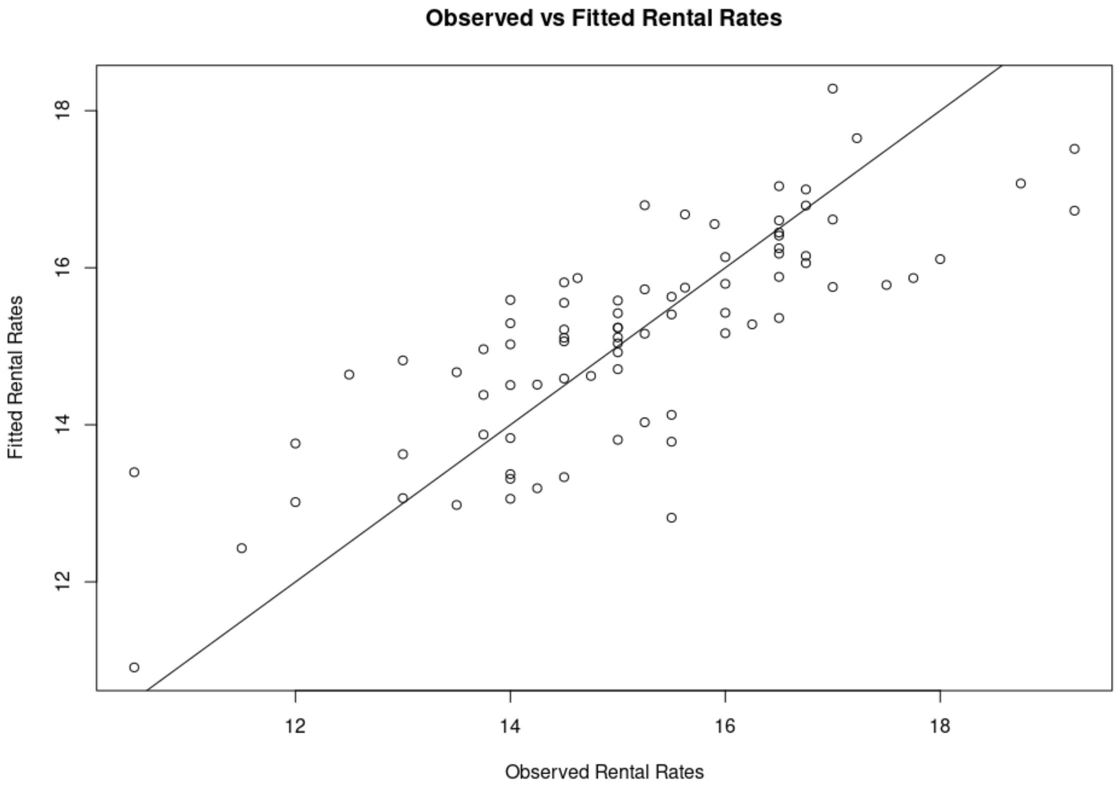
Answer:

R Code:



```
setwd("/home/swaroop/Downloads/Assignments/STP530/HW9")

data <- read.table("CH06PR18.txt", header=FALSE)
names(data) <- c("Y", "X1", "X2", "X3", "X4")

# a

data$x1_centered <- data$X1 - mean(data$X1)
data$x1_centered_squared <- data$x1_centered^2

model <- lm(Y ~ x1_centered + x1_centered_squared + X2 + X4, data=data)

plot(data$Y, fitted(model), xlab = "Observed Rental Rates", ylab = "Fitted Rental Rates", main = "Observed vs Fitted
Rental Rates")
abline(0, 1)  # Adds a 45-degree line to the plot
```

R Output:

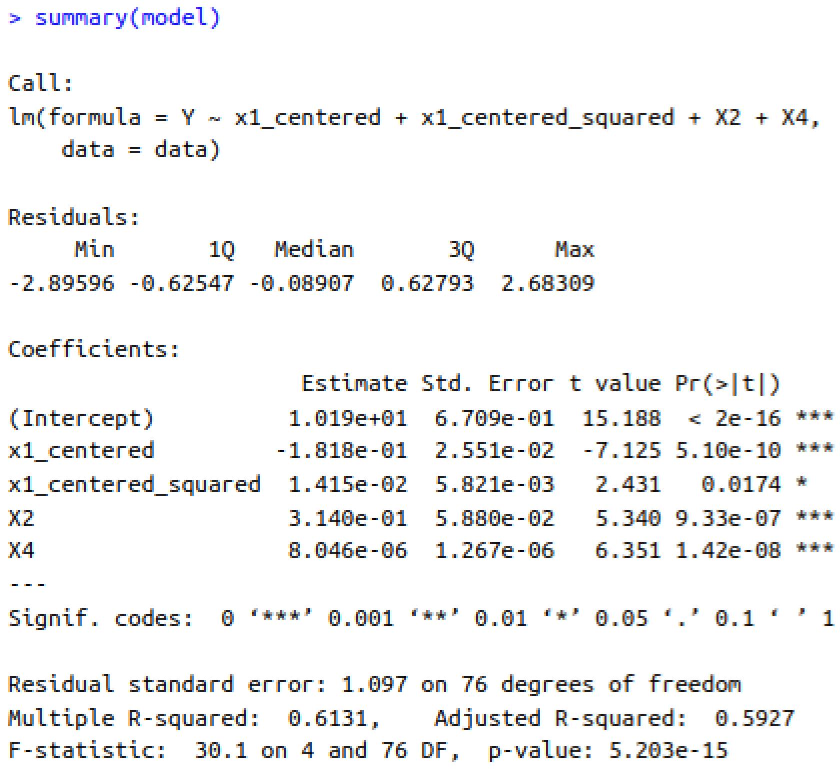


The scatter plot of observed versus fitted rental rates shows that most of the data points cluster around an imaginary line that would pass through the origin and follow a positive diagonal across the plot, which suggests a good fit. There is a clear positive correlation between the fitted values and the observed data, indicating that as the observed rental rates increase, the fitted values also increase correspondingly. This pattern suggests that the model is capturing the general trend in the data well.

b. Calculate Ra^2. What information does this measure provide?

Answer:

R Output:



```
> summary(model)

Call:
lm(formula = Y ~ x1_centered + x1_centered_squared + X2 + X4,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.89596 -0.62547 -0.08907  0.62793  2.68309 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          1.019e+01  6.709e-01  15.188  < 2e-16 ***
x1_centered         -1.818e-01  2.551e-02  -7.125 5.10e-10 ***
x1_centered_squared  1.415e-02  5.821e-03   2.431   0.0174 *  
X2                   3.140e-01  5.880e-02   5.340 9.33e-07 ***
X4                   8.046e-06  1.267e-06   6.351 1.42e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.097 on 76 degrees of freedom
Multiple R-squared:  0.6131,	Adjusted R-squared:  0.5927 
F-statistic:  30.1 on 4 and 76 DF,  p-value: 5.203e-15
```

In the given model, the adjusted R^2 of 0.5927 suggests that the model explains 59.27% of the variance in rental rates, after adjusting for the number of predictors. This value indicates a moderate level of explanatory power, showing that the model fits the data reasonably well, while also accounting for the fact that multiple predictors are being used.

c. Test whether or not the the square of centered property age (x1^2) can be dropped from the model; use alpha = .05. State the alternatives, decision rule and conclusion. What is the p-value of the test?
d. Estimate the mean rental rate when X1 = 8. X2 = 16 and X4 = 250,000. use a 95 percent confidence interval. Interpret your interval.

Answer:

Step 1: Assumptions:
The errors, epsilon , are assumed to be independent and identically distributed (i.i.d. following a normal distribution with mean 0 and constant variance sigma^2 .

Step 2: Hypotheses:

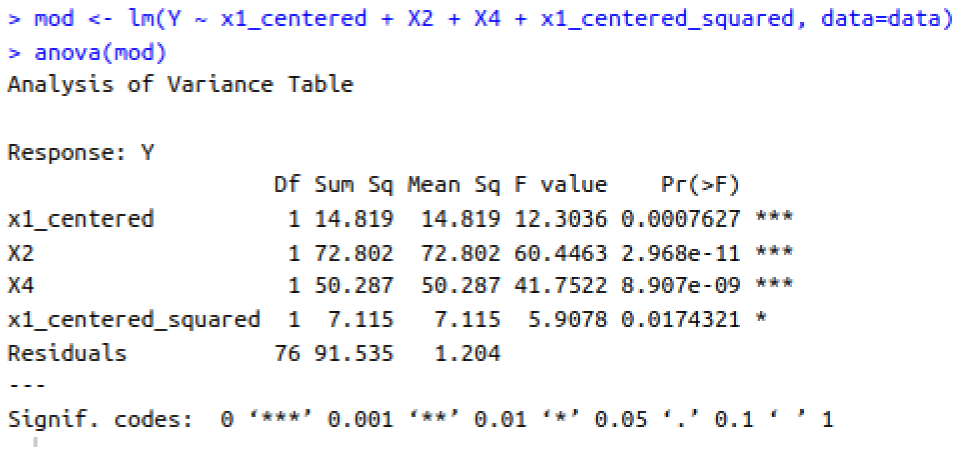
Full Model (includes beta_3 for x1^2): E(Y)=beta_0 + beta_1 * x1 + beta_11* x1^2 + beta_2 * X2 + beta_4 * X4

Reduced Model (excludes beta_11 for x1^2): E(Y)=beta_0 + beta_1 * x1 + beta_2 * X2 + beta_4 * X4

Hypotheses: H0 : beta_11=0; H1 : beta_11!=0

Step 3: Test-statistic: (All relevant quantities are directly available in the ANOVA table above.)

R Output:



```
> mod <- lm(Y ~ x1_centered + X2 + X4 + x1_centered_squared, data=data)
> anova(mod)
Analysis of Variance Table

Response: Y
                    Df Sum Sq Mean Sq F value    Pr(>F)    
x1_centered          1 14.819  14.819 12.3036 0.0007627 ***
X2                   1 72.802  72.802 60.4463 2.968e-11 ***
X4                   1 50.287  50.287 41.7522 8.907e-09 ***
x1_centered_squared  1  7.115   7.115  5.9078 0.0174321 *  
Residuals           76 91.535   1.204                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

F* = 5.9078

Step 4: P-value:

From anova table p-value = 0.0174

Step 5: Conclusion:

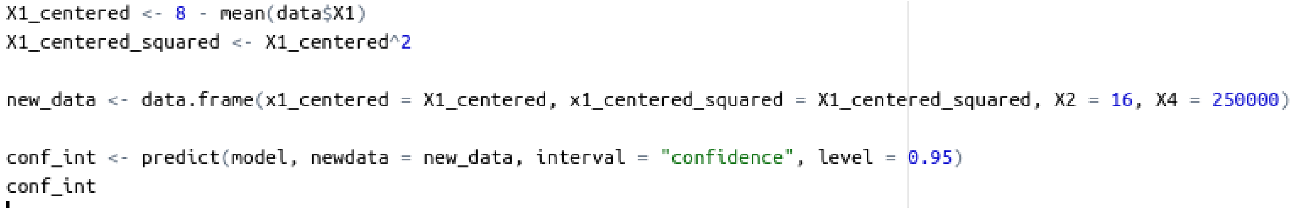
Note that the level of significance required by this problem is α = .05. Because the P-value (0.01743) is less than α = 0.05, we reject H_0 at the significance level of 0.05. This means that the full model fits the data significantly better than the reduced model, so we should not drop x1_centered_squared from the model. Putting this in context, with the other predictors already in the model, the squared term of the centered x1 variable does enhance the prediction of the response variable Y.

d. Estimate the mean rental rate when X1 = 8. X2 = 16 and X4 = 250,000. use a 95 percent confidence interval. Interpret your interval.

Answer:

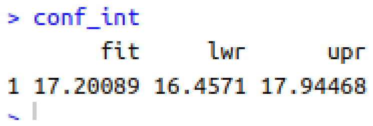
Transformed Model:

R Code:



```
X1_centered <- 8 - mean(data$X1)
X1_centered_squared <- X1_centered^2

new_data <- data.frame(x1_centered = X1_centered, x1_centered_squared = X1_centered_squared, X2 = 16, X4 = 250000)

conf_int <- predict(model, newdata = new_data, interval = "confidence", level = 0.95)
conf_int
```

R Output:



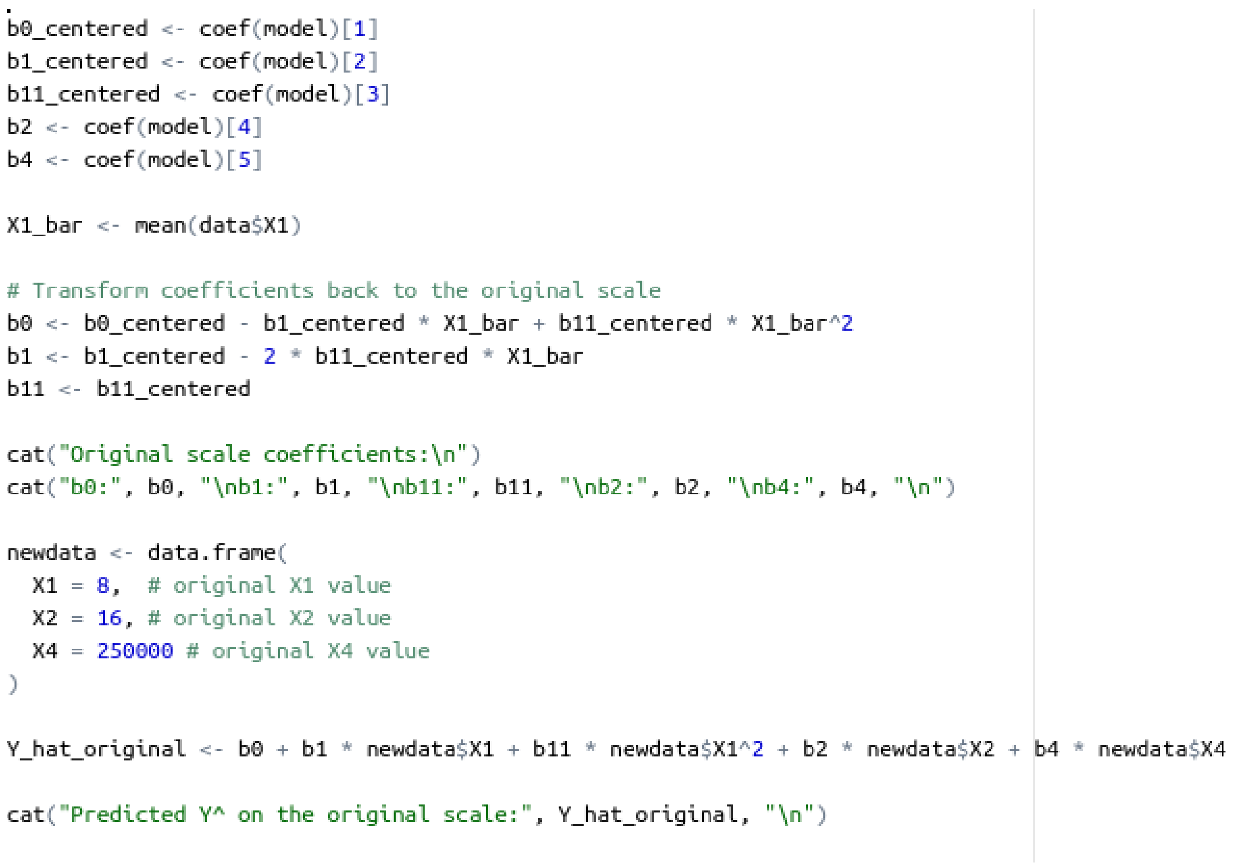
```
> conf_int
       fit     lwr      upr
1 17.20089 16.4571 17.94468
```

Interpretation:

Confidence interval for E{Y}: We are 95% sure that the mean rental rate for all observations with the given values of X1, X2, and X4 (8, 16, and 250,000 respectively) falls between 16.4571 and 17.94468.

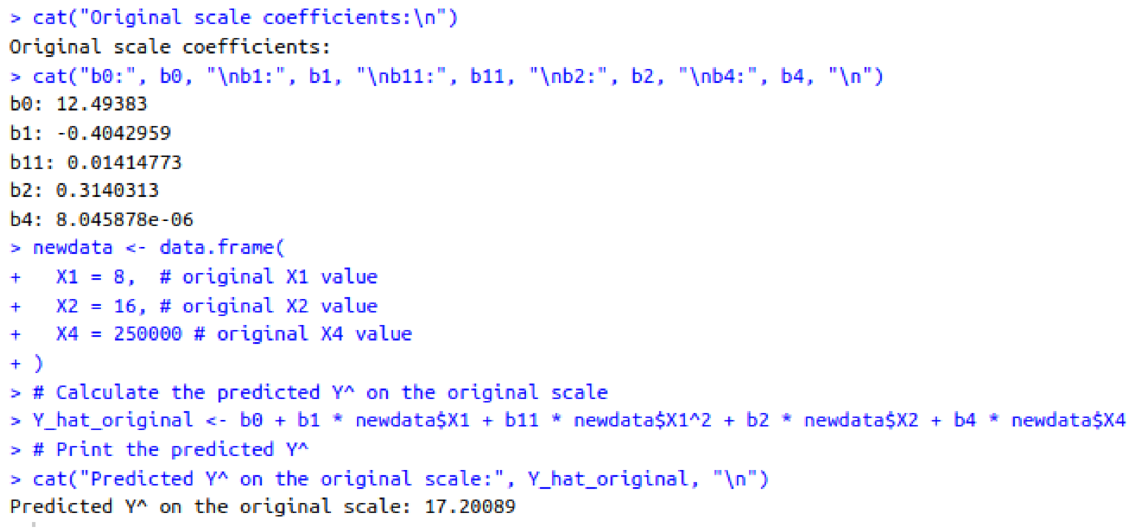
Original Scale:

R Code:



```
b0_centered <- coef(model)[1]
b1_centered <- coef(model)[2]
b11_centered <- coef(model)[3]
b2 <- coef(model)[4]
b4 <- coef(model)[5]

X1_bar <- mean(data$X1)

# Transform coefficients back to the original scale
b0 <- b0_centered - b1_centered * X1_bar + b11_centered * X1_bar^2
b1 <- b1_centered - 2 * b11_centered * X1_bar
b11 <- b11_centered

cat("Original scale coefficients:\n")
cat("b0:", b0, "\nb1:", b1, "\nb11:", b11, "\nb2:", b2, "\nb4:", b4, "\n")

newdata <- data.frame(
  X1 = 8,  # original X1 value
  X2 = 16, # original X2 value
  X4 = 250000 # original X4 value
)

Y_hat_original <- b0 + b1 * newdata$X1 + b11 * newdata$X1^2 + b2 * newdata$X2 + b4 * newdata$X4

cat("Predicted Y^ on the original scale:", Y_hat_original, "\n")
```

R Output:

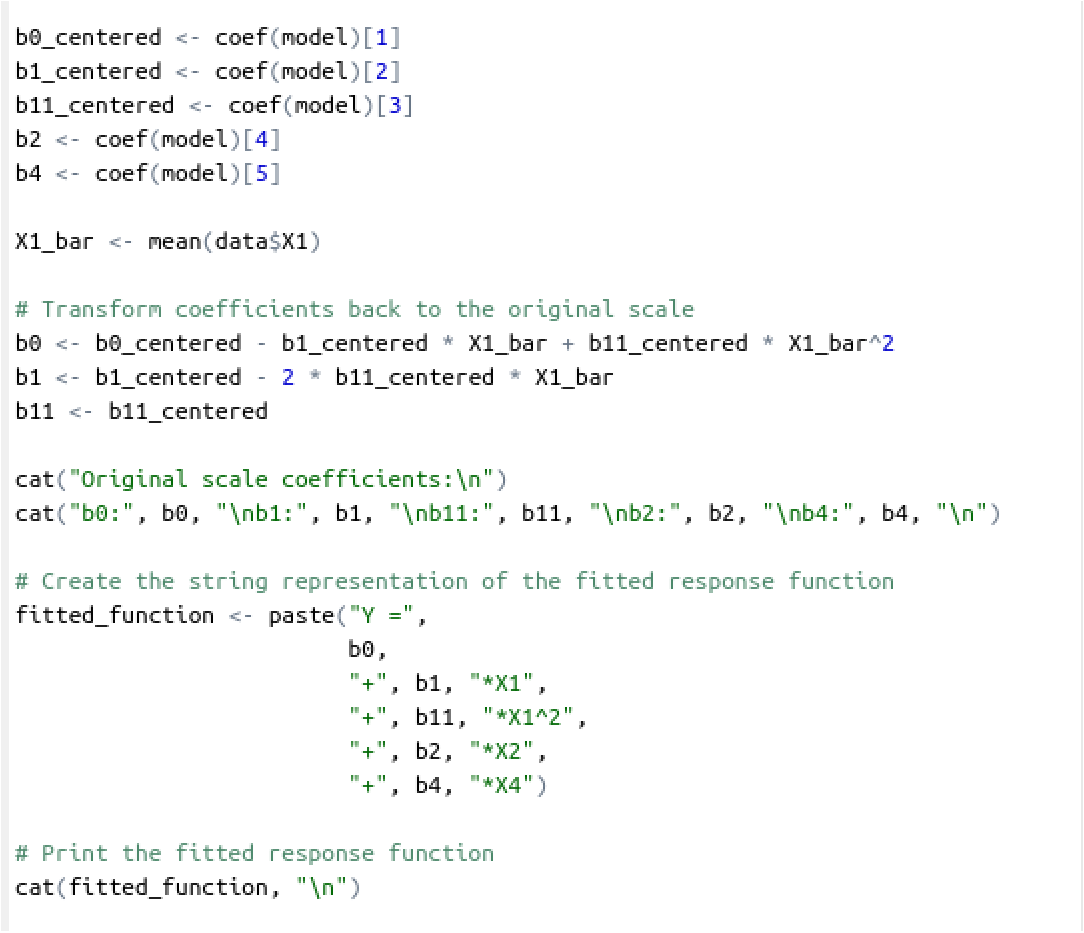


```
> cat("Original scale coefficients:\n")
Original scale coefficients:
> cat("b0:", b0, "\nb1:", b1, "\nb11:", b11, "\nb2:", b2, "\nb4:", b4, "\n")
b0: 12.49383
b1: -0.4042959
b11: 0.01414773
b2: 0.3140313
b4: 8.045878e-06
> newdata <- data.frame(
+   X1 = 8,  # original X1 value
+   X2 = 16, # original X2 value
+   X4 = 250000 # original X4 value
+ )
> # Calculate the predicted Y^ on the original scale
> Y_hat_original <- b0 + b1 * newdata$X1 + b11 * newdata$X1^2 + b2 * newdata$X2 + b4 * newdata$X4
> # Print the predicted Y^
> cat("Predicted Y^ on the original scale:", Y_hat_original, "\n")
Predicted Y^ on the original scale: 17.20089
```

e. Express the fitted response function obtained in part(a) in the original X variables.

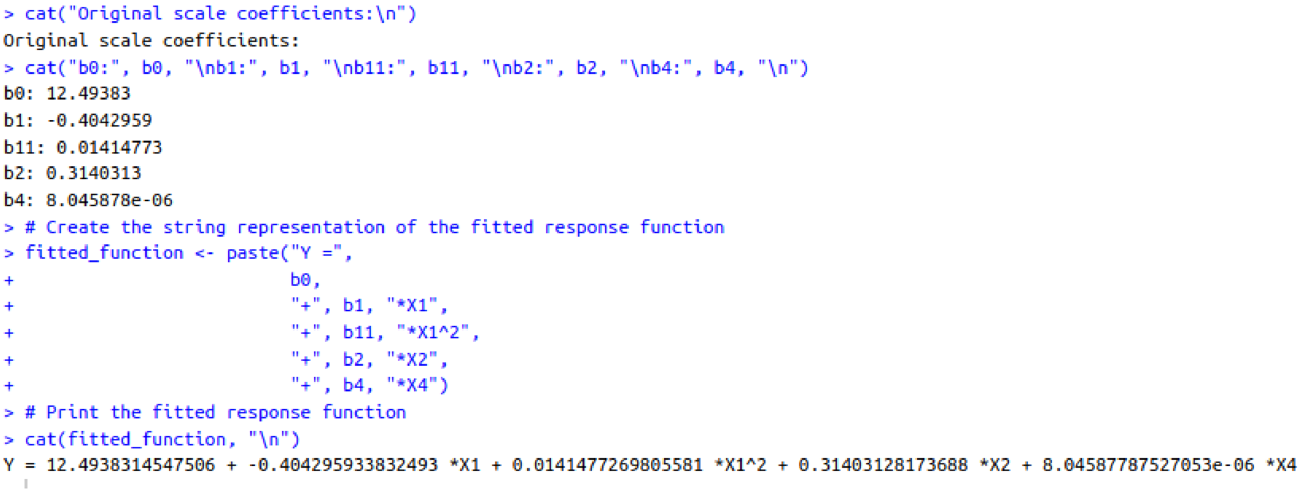
Answer:

R Code:



```
b0_centered <- coef(model)[1]
b1_centered <- coef(model)[2]
b11_centered <- coef(model)[3]
b2 <- coef(model)[4]
b4 <- coef(model)[5]

X1_bar <- mean(data$X1)

# Transform coefficients back to the original scale
b0 <- b0_centered - b1_centered * X1_bar + b11_centered * X1_bar^2
b1 <- b1_centered - 2 * b11_centered * X1_bar
b11 <- b11_centered

cat("Original scale coefficients:\n")
cat("b0:", b0, "\nb1:", b1, "\nb11:", b11, "\nb2:", b2, "\nb4:", b4, "\n")

# Create the string representation of the fitted response function
fitted_function <- paste("Y =",
                         b0,
                         "+", b1, "*X1",
                         "+", b11, "*X1^2",
                         "+", b2, "*X2",
                         "+", b4, "*X4")

# Print the fitted response function
cat(fitted_function, "\n")
```

R Output:



```
> cat("Original scale coefficients:\n")
Original scale coefficients:
> cat("b0:", b0, "\nb1:", b1, "\nb11:", b11, "\nb2:", b2, "\nb4:", b4, "\n")
b0: 12.49383
b1: -0.4042959
b11: 0.01414773
b2: 0.3140313
b4: 8.045878e-06
> # Create the string representation of the fitted response function
> fitted_function <- paste("Y =",
+                          b0,
+                          "+", b1, "*X1",
+                          "+", b11, "*X1^2",
+                          "+", b2, "*X2",
+                          "+", b4, "*X4")
> # Print the fitted response function
> cat(fitted_function, "\n")
Y = 12.4938314547506 + -0.404295933832493 *X1 + 0.0141477269805581 *X1^2 + 0.31403128173688 *X2 + 8.04587787527053e-06 *X4
```