Contents

Chapter 4

Multiple Regression Models

In this chapter, we explain regression models that are more complex than the first-order (straight-line) model with one explanatory variable. That is because in most practical applications of regression analysis we incorporate other potentially important independent variables into the model to make accurate predictions. These models that include more than one independent variable are called **multiple regression models**.

4.1 General Form of a Multiple Regression Model

General Form of the Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

The dependent variable y is now written as a function of k independent variables $x_1, x_2, ..., x_k$.

The random-error term is added to make the model probabilistic rather than deterministic.

Deterministic Portion of the Model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

The value of the coefficient β_i determines the contribution of the independent variable x_i, and $\beta_0$is the y-intercept. The coefficients $\beta_0, \beta_1, ..., \beta_k$are usually unknown because they represent population parameters.

$x_1, x_2, ..., x_k$, the independent variables, can be functions of variables, as long as the functions do not contain unknown parameters.

Note:

The symbols $x_1, x_2, ..., x_k$ may represent higher order terms for quantitative predictors (e.g., $x_2 = x_1^2$) or terms that represent qualitative predictors.

Finally, we use the same steps as we mentioned in chapter 3 to use the fitted model to estimate the mean value of y or to predict a particular value of y for given values of the independent variables, and to make other inferences.

4.2 Model Assumptions

Assumptions about Random Error ϵ

For any given set of values of $x_1, x_2, ..., x_k$ the random error ϵ has a probability distribution with the following properties:

1. The mean is equal to 0.
2. The variance is equal to σ^2
3. The probability distribution is a normal distribution.
4. Random errors are independent (in a probabilistic sense).

Simple Linear Regression in Matrix Form

$$Y = X\beta + \epsilon$$

where:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & ... & x_{1k} \\ 1 & x_{21} & ... & x_{2k} \\ . & . & & \\ . & . & & \\ . & . & & \\ 1 & x_{n1} & ... & x_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

$\hat{\beta} = (X'X)^{-1}X'Y$, where $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ . \\ . \\ . \\ \hat{\beta}_k \end{bmatrix}$

4.3 A First-Order Model with Quantitative Predictors

A First-Order Model in K Quantitative Independent Variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

where $x_1, x_2, ..., x_k$ are all quantitative variables that are not functions of other independent variables.

Note: β_i represents the slope of the line relating y to x_i when all the other x's are held fixed.

Interpretation of β_i when the corresponding independent variable is quantitative:

The β_i parameter in the first-order model above has similar interpretations that we had in chapter 3 considering the values of the remaining independent variables are held to be fixed.

4.4 Fitting the Model: The Method of Least Squares

The method of fitting multiple regression models is similar to the method of least squares.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_1 x_k$$

with minimum value of SSE. $SSE = \sum (y - \hat{y})^2$

When $k + 1$ simultaneous linear equations must be solved to find $k + 1$ estimated coefficients, $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_1$.

Example 4.4.1 *Refer to Exercise 1.62,* ***Cooling method for gas turbines****. During periods of high electricity demand, especially during the hot summer months, the power output from a gas turbine engine can drop dramatically. One way to counter this drop in power is by cooling the inlet air to the gas turbine. An increasingly popular cooling method uses high pressure inlet fogging. The performance of a sample of 67 gas turbines augmented with high pressure inlet fogging was investigated in the Journal of Engineering for Gas Turbines and Power (January 2005). Data set: GASTURBINE*

Variables:

HEATRATE: Heat rate (kilojoules per kilowatt per hour).

RPM: Cycle Speed: (revolutions per minute) of the engine.

INLET.TEMP: Inlet Temperature, (°C).

EXH.TEMP: Exhaust Gas Temperature, (°C).

CPRATIO: Cycle Pressure Ratio.

AIRFLOW: Air Mass Flow Rate (kilograms per second).

a) Use scatter-plot-matrix to plot the sample data. Interpret the plots.

First, we use a scatter-plot-matrix to visualize the relationships between each pair of variables and how the response variable, HEATRATE, relates to all five predictors simultaneously.

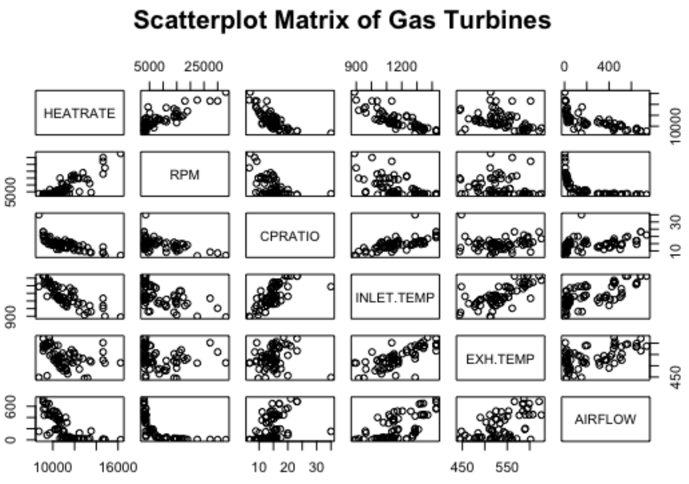
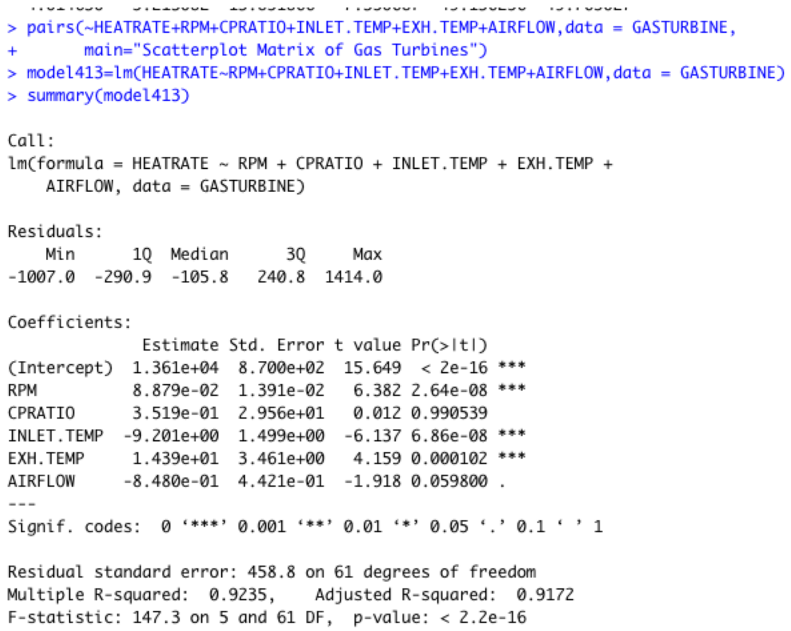


Figure 4.1: Scatter plot matrix of the response and the five predictor variables

b) Write a first-order model for heat rate (y) as a function of speed, inlet temperature, exhaust temperature, cycle pressure ratio, and air flow rate.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_5 x_5$$

c) Fit the model to the data using the method of least squares.



```
> pairs(~HEATRATE+RPM+CPRATIO+INLET.TEMP+EXH.TEMP+AIRFLOW,data = GASTURBINE,
+       main="Scatterplot Matrix of Gas Turbines")
> model413=lm(HEATRATE~RPM+CPRATIO+INLET.TEMP+EXH.TEMP+AIRFLOW,data = GASTURBINE)
> summary(model413)

Call:
lm(formula = HEATRATE ~ RPM + CPRATIO + INLET.TEMP + EXH.TEMP +
    AIRFLOW, data = GASTURBINE)

Residuals:
    Min      1Q  Median      3Q     Max
-1007.0  -290.9  -105.8   240.8  1414.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.361e+04  8.700e+02  15.649  < 2e-16 ***
RPM          8.879e-02  1.391e-02   6.382 2.64e-08 ***
CPRATIO      3.519e-01  2.956e+01   0.012 0.990539
INLET.TEMP  -9.201e+00  1.499e+00  -6.137 6.86e-08 ***
EXH.TEMP     1.439e+01  3.461e+00   4.159 0.000102 ***
AIRFLOW     -8.480e-01  4.421e-01  -1.918 0.059800 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 458.8 on 61 degrees of freedom
Multiple R-squared:  0.9235,	Adjusted R-squared:  0.9172
F-statistic: 147.3 on 5 and 61 DF,  p-value: < 2.2e-16
```

The least squares prediction equation according to the Rstudio output is:

$$\hat{y} = 13610 + 0.08879x_1 + 0.3519x_2 - 9.201x_3 + 14.39x_4 - 0.848x_5$$

when $x_1 =$ RPM, $x_2 =$ CPRATIO, $x_3 =$ INLET.TEMP, $x_4 =$ EXH.TEMP, and $x_5 =$ AIRFLOW

d) Give practical interpretations of the β estimates.

$\hat{\beta}_1 = 0.08879$ means the mean heat rate is estimated to increase by 0.08879 kilojoules per kilowatt per hour for every 1 revolution per minute increase in speed of the engine. when other indepndent variables are held fixed.

4.5 Estimation of σ^2, the Variance of ϵ

σ^2 is the variance of the random error, ϵ.

If $\sigma^2 = 0 \rightarrow \epsilon = 0 \longrightarrow \hat{y} = E(y)$

However, larger σ^2 means larger random error, ϵ and larger deviations between $\hat{y}$ and $E(y)$ and larger error in estimating the model parameters and larger error in predicting a value of y for a specific set of values of $x_1, x_2, ..., x_k$.

Estimator of σ^2 for a Multiple-Regression Model with k Independent Variables

$$s^2 = \frac{SSE}{\text{Degrees of freedom for error}} = \frac{SSE}{n - \text{Number of estimated} \quad \beta \quad parameters} = \frac{SSE}{n-(k+1)} \quad \text{(Mean Square Error)}$$

Example 4.5.1 *Refer to Example 4.4.1, find the model standard deviation, s, and interpret its value.*

$s = 458.8$

Thus, we expect the model to provide predictions of heat rates to within about $\pm 2s = \pm 2(458.8) = \pm 917.6$ kilojoules per kilowatt per hour. Or, about 95% of sample heat rates fall within 917.6 kilojoules per kilowatt per hour of their predicted values using the model.

4.6 Testing the Utility of a Model: The Analysis of Variance F-Test

To test the utility of the model or to detrmine whether the model is adequate for predicting y, we conduct a **global test**. Because conducting a series of t-tests to determine whether the independent variables are contributing to the predictive relationship inflates the overall Type I error and may include a large number of insignificant variables and exclude some useful ones.

Testing the Global Usefulness of the Model: The Analysis-of-Variance F-Test

$H_0 : \beta_1 = \beta_2 ... = \beta_k = 0$

$H_a :$ At least one$\beta_i \neq 0$

Test statistic: $F_c = \frac{\frac{SSyy - SSE}{k}}{\frac{SSE}{n-(k+1)}} = \frac{\frac{R^2}{k}}{\frac{(1-R^2)}{n-(k+1)}} = \frac{\text{Mean square (Model)}}{\text{Mean square (Error)}}$

where n is the sample size and k is the number of terms in the model.

Rejection region: $F > F_\alpha$

p-value: $P(F > F_c)$

where the F-distribution has k numerator degrees of freedom and $[n - (k + 1)]$ denominator degrees of freedom.

Assumptions: The standard regression assumptions about the random error component (Section 4.2).

Example 4.6.1 *Refer to Example 4.4.1, Is the overall model statistically useful at predicting heat rate (y)? Test using* $\alpha = .01$.

$F = 147.3 \longrightarrow p\text{-value} << 2.2e - 16 \approx 0 < \alpha = 0.01$

Decision: We reject H_0.

Conclusion: At 1% level of significance, The date provide strong evidence to conclude that the model is useful in predicting the heat rate.

Note: Statistically "useful" does not necessarily mean "best." We might find another model that is more useful in terms of providing more reliable estimates and predictions. The model must pass the F-test, but "pass" doesn't mean "best".

Example 4.6.2 *Refer to Example 4.4.1, do you recommend using the model in practice? Explain.*

No. Since ....

$ANOVA$ Table for a Multiple Regression Model with k Independent Vaiables

$ANOVA$ Table for a Multiple Regression Model with k Independent Vaiables				
Source	Degrees of Freedom df	Sum of Squares SS	Mean Sqare MS	F
Regression	k	SSR	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Error	$n-k-1$	SSE	$MSE = \frac{SSE}{n-k-1}$	
Total	$n-1$	SST		

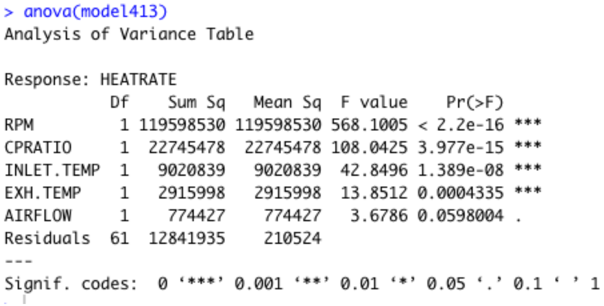
When:

$$SST = SSR + SSE$$

$$SST = SS_{yy} = \sum (y_i - \bar{y})^2$$

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$



```
> anova(model413)
Analysis of Variance Table

Response: HEATRATE
           Df    Sum Sq   Mean Sq  F value    Pr(>F)    
RPM         1 119598530 119598530 568.1005 < 2.2e-16 ***
CPRATIO     1  22745478  22745478 108.0425 3.977e-15 ***
INLET.TEMP  1   9020839   9020839  42.8496 1.389e-08 ***
EXH.TEMP    1   2915998   2915998  13.8512 0.0004335 ***
AIRFLOW     1    774427    774427   3.6786 0.0598004 .  
Residuals  61  12841935    210524                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4.7 Inferences About the Individual β Parameters

Test of an Individual Parameter Coefficient in the Multiple Regression Model

Lower-Tailed Test	Upper-Tailed Test	Two-Tailed Test
$H_0 : \beta_1 = 0$	$H_0 : \beta_1 = 0$	$H_0 : \beta_1 = 0$
$H_a : \beta_1 < 0$	$H_a : \beta_1 > 0$	$H_a : \beta_1 \neq 0$

Test statistic:

$$t_c = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$$

Rejection Region:

$t < -t_\alpha$	$t > t_\alpha$	$\lvert t \rvert > t_{\frac{\alpha}{2}}$

where t_α and $t_{\frac{\alpha}{2}}$ are based on $n - (k+1)$ degrees of freedom and:

$n =$ Number of observations

$k + 1 =$ Number of β parameters in the model

Note: Most statistical software programs report two-tailed p-values on their output. To find the appropriate p-value for a one-tailed test, make the following adjustment to $P =$ two-tailed p-value:

Upper-Tailed Test $(H_a : \beta_1 > 0)$

$$p - value = \begin{cases} \frac{p}{2}, & \text{if } t > 0 \\ 1 - \frac{p}{2}, & \text{if } t < 0 \end{cases}$$

Lower-Tailed Test $(H_a : \beta_1 < 0)$

$$p - value = \begin{cases} \frac{p}{2}, & \text{if } t < 0 \\ 1 - \frac{p}{2}, & \text{if } t > 0 \end{cases}$$

where p is the p-value reported on the printout and t is the value of the test statistic.

Assumptions: See Section 4.2 for assumptions about the probability distribution for the random error component ϵ.

A 100$(1-\alpha)$% Confidence Interval for a β

$$\hat{\beta}_i \pm (t_{\frac{\alpha}{2}}) s_{\hat{\beta}_i}$$

where $t_{\frac{\alpha}{2}}$ is based on n - (k + 1) degrees of freedom.

$n =$ Number of observations

$k + 1 =$ Number of β parameters in the model

Example 4.7.1 *Refer to Example 4.4.1, test to determine whether AIRFLOW, x_5 (kilograms per second) is a useful linear predictor of HEATRATE, y (kilojoules per kilowatt per hour). using $\alpha = .05$.*

To determine if AIRFLOW is a useful linear predictor of HEATRATE, we test:

$H_0 : \beta_5 = 0$

$H_a : \beta_2 \neq 0$

Test statistic: $t_c = \frac{\hat{\beta}_5}{s_{\hat{\beta}_5}} = \frac{-0.848}{0.4421} = -1.918$

p-value $= 0.059800 > 0.05 \longrightarrow$ we fail to reject H_0.

At 5% level of significance, there is insufficient evidence to indicate that AIRFLOW is a useful linear predictor of HEATRATE.

Recommendation for Checking the Utility of a Multiple Regression Model

1. Conduct F- test to test the utility of the model. (Global Usefulness of the Model) If you rejected the H_0, proceed to step 2. Otherwise, fit another model with more independent variables or higher-order terms.

2. Conduct t-test(s) about Individual parameter coefficient(s) just for the most important ones. (Why?)

Question: What are the possible conclusions if we fail to reject $H_0 : \beta_i = 0$ in a first-order linear model for the purpose of determining which independent variables are useful for predicting y?

4.8 Multiple Coefficients of Determination: R^2 and R_a^2

The multiple coefficient of determination, R^2 , is defined as

$$R^2 = \frac{(SS_{yy} - SSE)}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}} = \frac{\text{Explained variability}}{\text{Total variability}}$$

and represents the proportion of the total sample variation in y that can be "explained" by the multiple-regression model.

$$0 \leq R^2 \leq 1$$

The high value of R^2, from Example 4.4.1, implies that using the independent variables in a first-order model explains 92.35% of the total sample variation (measured by $SSyy$) in HEATRATE y. Thus, R^2 is a sample statistic that tells how well the model fits the data and thereby represents a measure of the usefulness of the entire model.

Question: Does a high value of R^2 mean the model provides a good fit to all of the data points in the population?

R^2 increases by increasing the number of predictors in the model.

Since you will always obtain a perfect fit R^2 to a set of n data points if the model contains exactly n parameters, the adjusted multiple coefficient of determination, denoted R_a^2, is often reported.

Note that R_a^2= .9172, a value only slightly smaller than R^2.

$$R_a^2 = 1 - [\frac{(n-1)}{n-(k+1)}](\frac{SSE}{SS_{yy}})$$

or

$$R_a^2 = 1 - [\frac{(n-1)}{n-(k+1)}](1 - R^2)$$

Note: $R_a^2 < R^2$ and might be negative if the model fits poorly.

Note: R_a^2 and R^2 have similar interpretations. R_a^2 idicates how well a multiple regression model fits a set of data and also adjusts for the number of terms in a model and sample size.

Example 4.8.1 *Refer to Example 4.4.1, find the adjusted-R^2 value and interpret it.*

91.7% of the variation in HEATRATE can be explained by the fitted model.

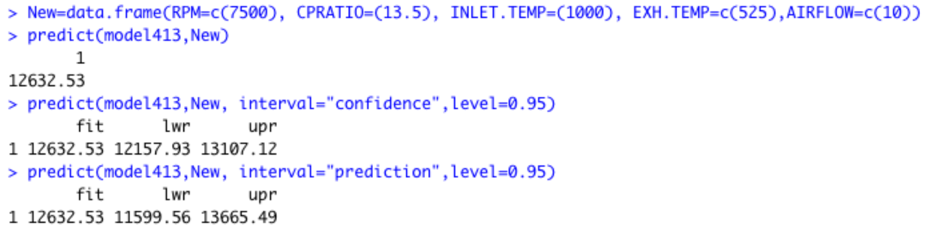
4.9 Using the Model for Estimation and Prediction

As we discussed in chapater 3, we would like to use the least square regression line for estimationg the mean value of y, $E(y)$, for some value of x and predict some future value of y using some value of x.

Example 4.9.1 *Refer to Example 4.4.1, use the RStudio output below and interprete the 95% confidence interval for $E(y)$ and 95% prediction interval for y in the words of the problem when:*

RPM= 7500, INLET-TEMP= 1000, EXH-TEMP= 525, CPRATIO= 13.5, AIRFLOW= 10

The predicted value of HEATRATE, $\hat{y}$ or the estimated mean value of HEAT RATE, $\hat{E(y)}$ for the observation above is 12632.53 (kilojoules per kilowatt per hour).



```
> New=data.frame(RPM=c(7500), CPRATIO=(13.5), INLET.TEMP=(1000), EXH.TEMP=c(525),AIRFLOW=c(10))
> predict(model413,New)
       1
12632.53
> predict(model413,New, interval="confidence",level=0.95)
       fit      lwr      upr
1 12632.53 12157.93 13107.12
> predict(model413,New, interval="prediction",level=0.95)
       fit      lwr      upr
1 12632.53 11599.56 13665.49
```

We are 95% confident to say that the mean heat rate level will range from 12157.93 and 13107.12 kilojoules per kilowatt per hour for **all possible gas turbins** with RPM= 7500, INLET-TEMP= 1000, EXH-TEMP= 525, CPRATIO= 13.5, AIRFLOW= 10.

We can also with 95% confident predict that the heat rate level will fall into the interval from 11599.56 and 13665.49 kilojoules per kilowatt per hour for **a gas turbin** with RPM= 7500, ILET-TEMP= 1000, EXH-TEMP= 525, CPRATIO= 13.5, AIRFLOW= 10.

4.10 An Interaction Model with Quantitative Predictors

In the straight-line model (simple linear regression), the slope of the line, β_1 represents the mean change in y for every 1-unit increase in x. Similar interpretations when the independent variables are quantitative can be applied in the first-order model by holding the values of the remaining independent variables fixed.

Suppose the first-order model is:

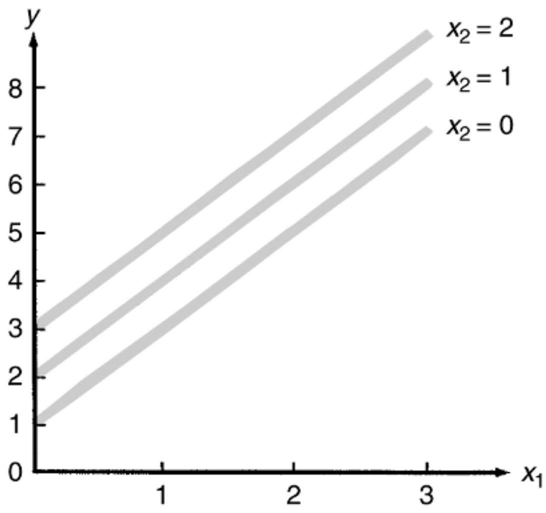
$$E(y) = 1 + 2x_1 + x_2$$

If $x_2 = 0 \longrightarrow E(y) = 1 + 2x_1$

If $x_2 = 1 \longrightarrow E(y) = 1 + 2x_1 + 1 = 2 + 2x_1$

Graphs of $E(y) = 1 + 2x_1 + x_2$ for $x_2 = 0, 1, 2$

The lines are parallel since the slopes of the three lines are all equal to $\beta_2 = 2$.



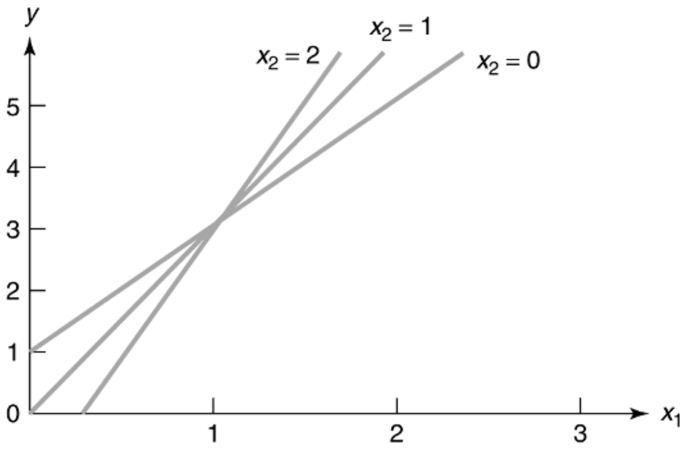
For all first-order models, $E(y)$ is linearly related to any one variable, x_i for the fixed values of the other variables with slope of β_i. By repeating the process for other values of the fixed independent variables, we have a set of parallel straight lines that indicates the effect of x_i on $E(y)$ is independent of all the other independent variables in the model and is measured by the slope β_i.

Threfore, we can conclude that the relationship between $E(y)$ and any one independent variable does not depend on the values of the other independent variables in the model. Otherwise the first-order model is not appropriate for predicting y and we need to take into account this dependence. The new model includes the **cross products** of two or more independent variables.

For example:

$$E(y) = 1 + 2x_1 - x_2 + x_1 x_2$$

Suppose: $x_2 = 0, 1, 2$



The lines are not parallel and do not have the same slope.

$x_2 = 0 \longrightarrow E(y) = 1 + 2x_1,\ \beta_1 = 2$

$x_2 = 1 \longrightarrow E(y) = 3x_1,\ \beta_1 = 3$

$x_2 = 2 \longrightarrow E(y) = -1 + 4x_1,\ \beta_1 = 4$

Note that the effect on $E(y)$ of a change in x_1 depends on the value of x_2 since x_1 and x_2 **interact**. The cross-product term is called **interaction term** and the model is called an **interaction model**.

An Interaction Model Relating $E(y)$ to Two Quantitative Independent Variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

where

$(\beta_1 + \beta_3 x_2)$ represents the change in $E(y)$ for every 1-unit increase in x_1, holding x_2 fixed.

$(\beta_2 + \beta_3 x_1)$ represents the change in $E(y)$ for every 1-unit increase in x_2, holding x_1 fixed.

Note: Do not conduct t-tests on the β coefficients of the first-order terms (x_1 and x_2), if the interaction is important for the model since the interaction implies both x_1 and x_2 are important.

Example 4.10.1 *Refer to Example 4.4.1:*

a) Researchers hypothesize that the linear relationship between heat rate (y) and temperature (both inlet and exhaust) depends on air flow rate. Write a model for heat rate that incorporates the researchers' theories.

b) Use statistical software to fit the interaction model, part a, to the data in the GASTURBINE file. Give the least squares prediction equation.

c) Conduct a test (at $\alpha = .05$) to determine whether inlet temperature and air flow rate interact to effect heat rate.

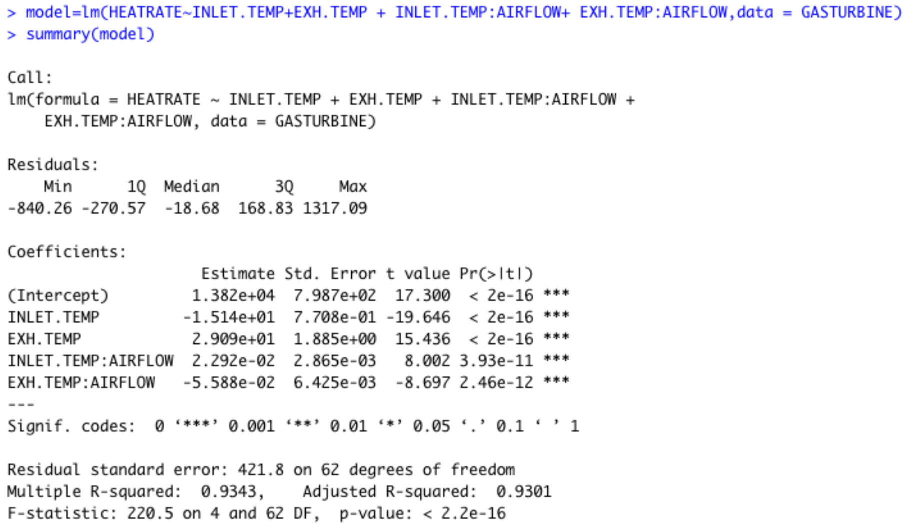
d) Conduct a test (at $\alpha = .05$) to determine whether exhaust temperature and air flow rate interact to effect heat rate.

e) Estimate the change in heat rate of a gas durbin when EXH-TEMP= 525, and AIRFLOW= 10 for each additional 1 (°C) of Inlet Temperature,.

a)

$$E(y) = \beta_0 + \beta_3 x_3 + \beta_4 x_4 + \beta_6 x_3 x_5 + \beta_7 x_4 x_5$$

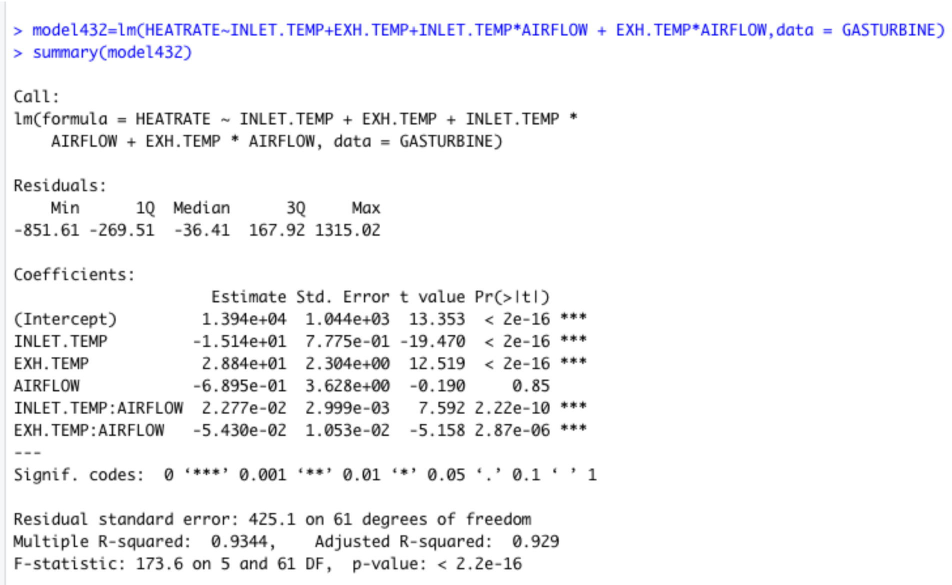
when x_1 = RPM, x_2 = CPRATIO, x_3 = INLET.TEMP, x_4 = EXH.TEMP, and x_5 = AIRFLOW b)



```
> model=lm(HEATRATE~INLET.TEMP+EXH.TEMP + INLET.TEMP:AIRFLOW+ EXH.TEMP:AIRFLOW,data = GASTURBINE)
> summary(model)

Call:
lm(formula = HEATRATE ~ INLET.TEMP + EXH.TEMP + INLET.TEMP:AIRFLOW + 
    EXH.TEMP:AIRFLOW, data = GASTURBINE)

Residuals:
    Min      1Q  Median      3Q     Max 
-840.26 -270.57  -18.68  168.83 1317.09 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         1.382e+04  7.987e+02  17.300  < 2e-16 ***
INLET.TEMP         -1.514e+01  7.708e-01 -19.646  < 2e-16 ***
EXH.TEMP            2.909e+01  1.885e+00  15.436  < 2e-16 ***
INLET.TEMP:AIRFLOW  2.292e-02  2.865e-03   8.002 3.93e-11 ***
EXH.TEMP:AIRFLOW   -5.588e-02  6.425e-03  -8.697 2.46e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 421.8 on 62 degrees of freedom
Multiple R-squared:  0.9343,	Adjusted R-squared:  0.9301 
F-statistic: 220.5 on 4 and 62 DF,  p-value: < 2.2e-16
```

$$\hat{y} = 13820 - 15.14x_3 + 29.09x_4 + 0.02292x_3x_5 - 0.05588x_4x_5$$

Question: Include AIRFLOW and see the difference.



```
> model432=lm(HEATRATE~INLET.TEMP+EXH.TEMP+INLET.TEMP*AIRFLOW + EXH.TEMP*AIRFLOW,data = GASTURBINE)
> summary(model432)

Call:
lm(formula = HEATRATE ~ INLET.TEMP + EXH.TEMP + INLET.TEMP * 
    AIRFLOW + EXH.TEMP * AIRFLOW, data = GASTURBINE)

Residuals:
    Min      1Q  Median      3Q     Max 
-851.61 -269.51  -36.41  167.92 1315.02 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         1.394e+04  1.044e+03  13.353  < 2e-16 ***
INLET.TEMP         -1.514e+01  7.775e-01 -19.470  < 2e-16 ***
EXH.TEMP            2.884e+01  2.304e+00  12.519  < 2e-16 ***
AIRFLOW            -6.895e-01  3.628e+00  -0.190     0.85    
INLET.TEMP:AIRFLOW  2.277e-02  2.999e-03   7.592 2.22e-10 ***
EXH.TEMP:AIRFLOW   -5.430e-02  1.053e-02  -5.158 2.87e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 425.1 on 61 degrees of freedom
Multiple R-squared:  0.9344,	Adjusted R-squared:  0.929 
F-statistic: 173.6 on 5 and 61 DF,  p-value: < 2.2e-16
```

c)

$H_0 : \beta_6 = 0$

$H_a : \beta_6 \neq 0$

p-value $\approx 0 < 0.05 \longrightarrow$ we reject H_0.

At 5% level of significance, there is sufficient evidence to conclude that the interaction between INLET-TEMP and AIRFLOW is significant at predicting HEATRATE.

d)

$H_0 : \beta_7 = 0$

$H_a : \beta_7 \neq 0$

p-value $\approx 0 < 0.05 \longrightarrow$ we reject H_0.

At 5% level of significance, there is sufficient evidence to conclude that the interaction between EXH.TEMP and AIRFLOW is significant at predicting HEATRATE.

e) To estimate the change in heat rate of a gas durbin when EXH-TEMP$=525$, and AIRFLOW$=10$ for each additional 1 ($°C$) of Inlet Temperature, x_3, we need to estimate the slope of the line relating y to x_3 when EXH-TEMP$=525$, and AIRFLOW$=10$.

4.11 A Quadratic (Second-Order) Model with a Quantitative Predictor

In this section, we talk about having a curvature in the relationship between $E(y)$ and independent variable(s).

When we include an x^2, the model is called **second-order model**.

Let's consider a second-order model with only one independent variable, x which is called the **quadratic model**.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

The term involving x^2, called a **quadratic term** or second-order term.

Consider the following graphs:

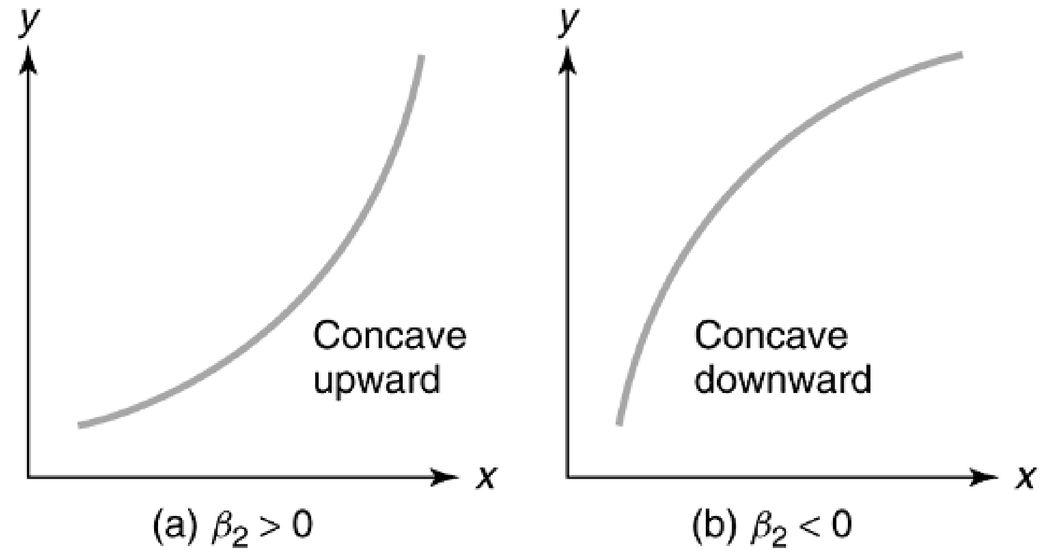


Figure 4.2: Graphs for two quadratic models

A Quadratic (Second-Order) Model in a Single Quantitative Independent Variable

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

where β_0 is the y-intercept of the curve

β_1 is a shift parameter

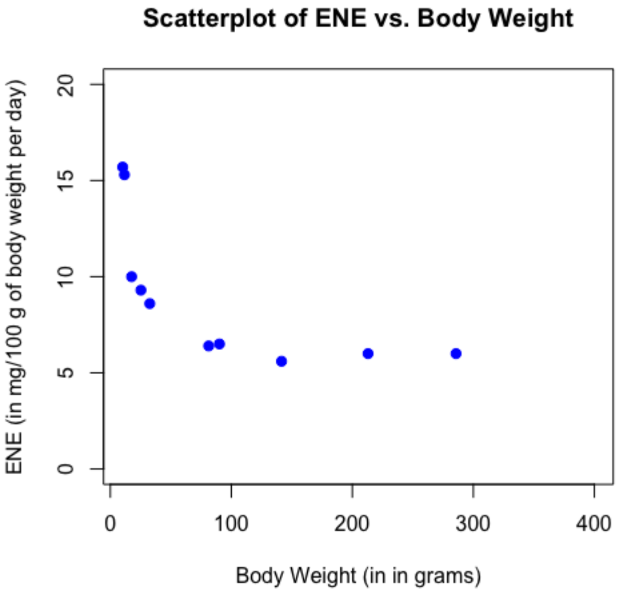
β_2 is the rate of curvature

Note:

- Interpreting of the estimated coefficients in a quadratic model must be undertaken cautiously.
- β_1, or the estimated coefficient of x does not represent a slope in the presence of the quadratic term x^2 and does not have a meaningful interpretation in the quadratic model.
- The sign of β_2, the estimated coefficient of the quadratic term, x^2, indicates whether the curve is concave downward (mound-shaped) or concave upward (bowl-shaped).

Example 4.11.1 *Refer to Example 4.37 from the textbook. A study of the variables that affect endogenous nitrogen excretion (ENE) in carp raised in Japan. Carps were divided into groups according to body weight and each group of 2-15 fish placed in a separate tank. The data set CARP shows the amount of ENE (in milligrams per 100 grams of body weight per day) measured in each tank after 20 days of feeding them a protein- free diet three times daily and the mean body weight (in grams) for each carp group.*

a) Graph the data in ascatter plot. Do you detect a pattern?



The scatterplot above shows that the amount of ENE appears to decrease in a curvilinear manner with the body weight.

b) The quadratic model $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$ was fit to the data using RStudio.

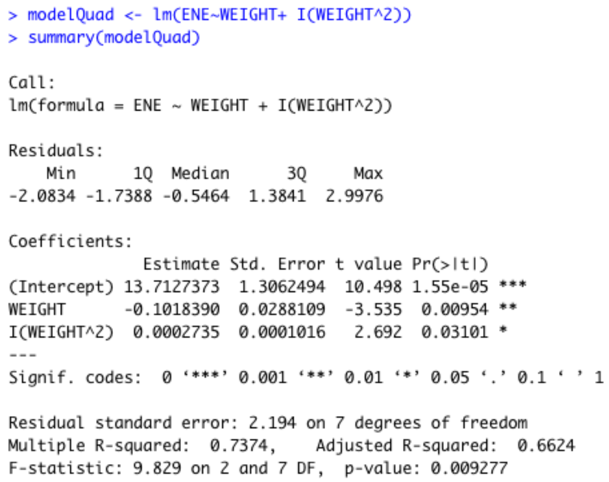
Conduct the test $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$ using $\alpha = .10$. Give the conclusion in the words of the problem.

$H_0 : \beta_2 = 0$

$H_a : \beta_2 \neq 0$

Decision: Since p-value $= 0.03101 < \alpha = .10$, we reject H_0.

Conclusion: At 10% level of significance, there is sufficient evidence to conclude that the body weight and ENT are quadratically related.



```
> modelQuad <- lm(ENE~WEIGHT+ I(WEIGHT^2))
> summary(modelQuad)

Call:
lm(formula = ENE ~ WEIGHT + I(WEIGHT^2))

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0834 -1.7388 -0.5464  1.3841  2.9976 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.7127373  1.3062494  10.498 1.55e-05 ***
WEIGHT      -0.1018390  0.0288109  -3.535  0.00954 ** 
I(WEIGHT^2)  0.0002735  0.0001016   2.692  0.03101 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.194 on 7 degrees of freedom
Multiple R-squared:  0.7374,	Adjusted R-squared:  0.6624 
F-statistic: 9.829 on 2 and 7 DF,  p-value: 0.009277
```

The RStudio output above shows the least squares estimates of the β parameters are

$\beta_0 = 13.7127373, \beta_1 = -0.1018390, \beta_2 = 0.0002735.$

Therefore, the equation that minimizes the SSE for the data is:

$\hat{y} = 13.713 - 0.102x + 0.0002x^2$

Note: The scatterplot above shows concave upward curvature in the relationship between ENE and Body Weight in the sample of 10 data points. To determine if this type of curvature exists in the population, we want to test:

$H_0 : \beta_2 = 0$

$H_a : \beta_2 > 0$

The test statistic for testing β_2 is 2.692 and the associated two-tailed

p-value is 0.03101.

Since this is a one-tailed test, the appropriate p-value is:

$p\text{-value} = \frac{0.03101}{2} = 0.015505 < \alpha = 0.1 \longrightarrow$, we reject H_0.

We have strong evidence of upward curvature in the population, that is, ENE decreases more slowly per 1 gram increase in body weight for carp groups with high body weight than for those with low body weight.

c) Graph the prediction equation and assess how well the model fits the data, both visually and numerically.

We can visualize that the graph provides a good fit to the data.

According to the RStudio output, $R_a^2 = 0.6624$ which implies about 66% of the sample variation in ENT (y) can be explained by the quadratic model (after adjusting for sample size and degrees of freedom). Note that Figure 4.3 implies that the estimated endogenous nitrogen excretion (ENE) is leveling off as the body weight levels increase between about 140 and 286 grams.

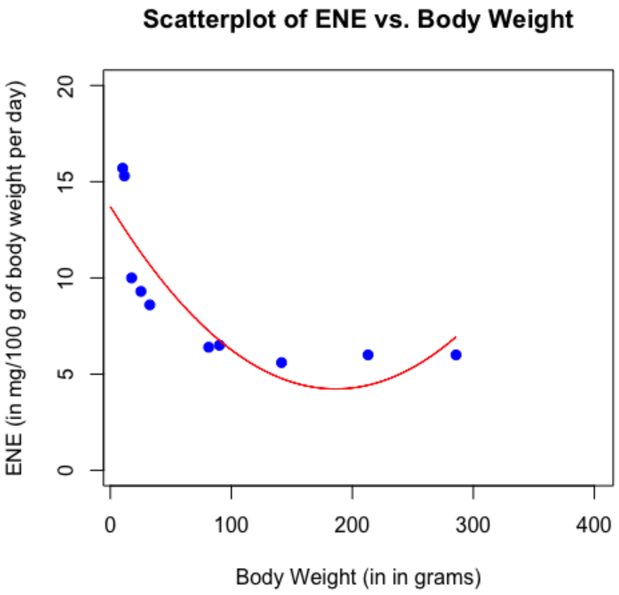


Figure 4.3: RStudio graph of least squares fit for the quadratic model

d) Is the overall model useful (at $\alpha = .01$) for predicting ENE?

Global F -test:

$H_0 : \beta_1 = \beta_2 = 0$

H_a : At least one of the above coefficients is nonzero

$F = 9.829$ with an associated p-value of 0.009277.

Since p-value of $0.009277 < 0.01$, we reject H_0 and conclude that the overall model is a useful predictor of ENE, y.

Measures of Overall Forecast Accuracy for n Forecasts

We can also use the following measures for comparing the overall accuracy of the regression models.

- Mean absolute percentage error: $MAPE = \frac{\sum_{i=1}^{n} |\frac{y_i - \hat{y}_i}{y_i}|}{n} \times 100$
- Mean absolute Error: $MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$
- Mean Square Error: $MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$
- Mean percentage error: $MPE = \frac{\sum_{i=1}^{n} \frac{y_i - \hat{y}_i}{y_i}}{n} \times 100$
- Root Mean Square Error: $RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$

4.12 More Complex Multiple Regression Models

In this section, we study some more advanced models. In previous sections we studied:

A First-Order Model Relating $E(y)$ to Five Quantitative x's

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_5 x_5$$

A Quadratic (Second-Order) Model Relating $E(y)$ to One Quantitative x

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

An Interaction Model Relating $E(y)$ to Two Quantitative x's

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Now we consider a a complete second-order model in two quantitative independent variables that has all of the terms in a first-order model and, in addition, the second-order terms involving cross-products (interaction terms) and squares of the independent variables.

A Complete Second-Order Model with Two Quantitative x's

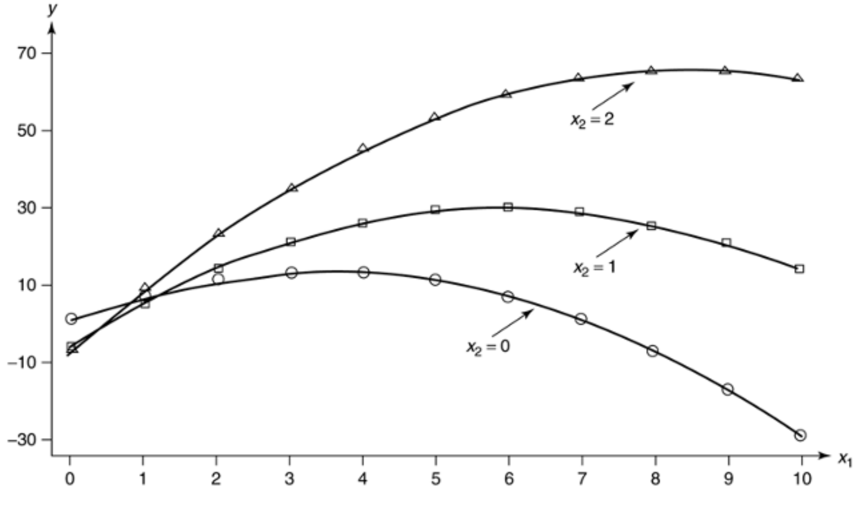
$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

If you have prior information (e.g. using scatterplot) that suggests there is moderate or very little curvature over the region in which the independent variables are measured, you could use the interaction model described previously.

Note:

You should have sufficient data to estimate all of the parameters in the second-order model.

Consider the following graph that shows the relationship between $E(y)$ and x_1 when $x_2 = 0, 1, 2$.



Graph of $E(y) = 1 + 7x_1 - 10x_2 + 5x_1x_2 - x_1^2 + 3x_2^2$

Example 4.12.1 *Consider Example 4.8 from the textbook. Data set: EXPRESS*

Dependent variable, y= Cost (in dollars)

Independent variables:

x_1 = Package Weight (in pounds)

x_2 = Distance Shipped (in miles)

a) Use scatter-plot-matrix to plot the sample data. Interpret the plots.

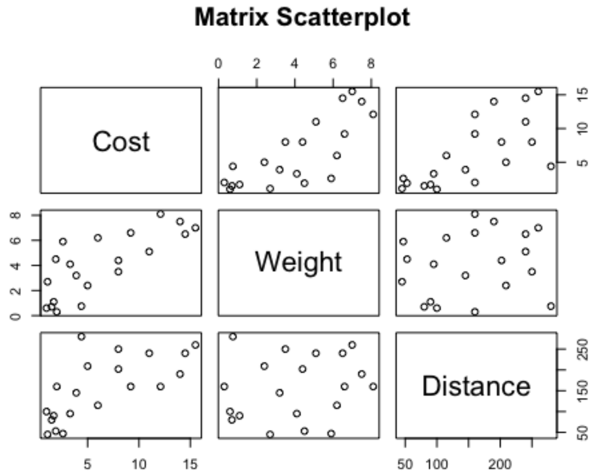
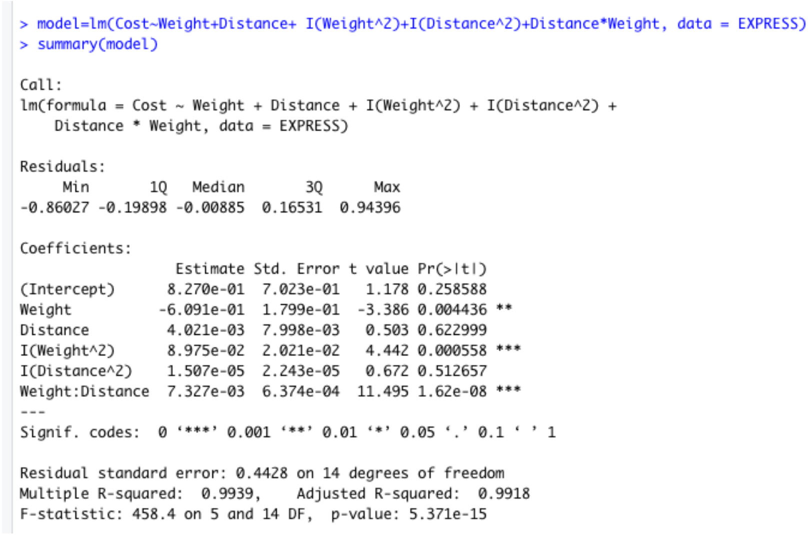


Figure 4.4: Matrix Scatterplot for data set EXPRESS

b) Give an appropriate linear model for the data.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \epsilon$$

c) Fit the model to the data and give the prediction equation.



```
> model=lm(Cost~Weight+Distance+ I(Weight^2)+I(Distance^2)+Distance*Weight, data = EXPRESS)
> summary(model)

Call:
lm(formula = Cost ~ Weight + Distance + I(Weight^2) + I(Distance^2) +
    Distance * Weight, data = EXPRESS)

Residuals:
     Min       1Q   Median       3Q      Max
-0.86027 -0.19898 -0.00885  0.16531  0.94396

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      8.270e-01  7.023e-01   1.178 0.258588
Weight          -6.091e-01  1.799e-01  -3.386 0.004436 **
Distance         4.021e-03  7.998e-03   0.503 0.622999
I(Weight^2)      8.975e-02  2.021e-02   4.442 0.000558 ***
I(Distance^2)    1.507e-05  2.243e-05   0.672 0.512657
Weight:Distance  7.327e-03  6.374e-04  11.495 1.62e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4428 on 14 degrees of freedom
Multiple R-squared:  0.9939,	Adjusted R-squared:  0.9918
F-statistic: 458.4 on 5 and 14 DF,  p-value: 5.371e-15
```

Figure 4.5: RStudio multiple regression output for Example 4.8

d) Find the value of s and interpret it.

$s = 0.4428$ which means 95% of of the sampled shipping cost values fall within $2s = .886$ of their respective predicted values.

e) Find the value of R_a^2 and interpret it.

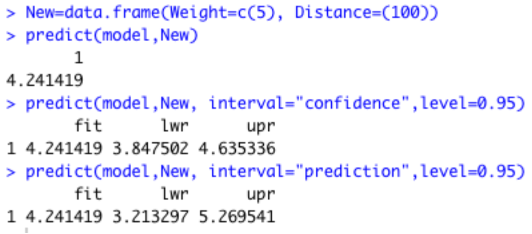
$R_a^2 = 0.9918$ which means after adjusting for sample size and the number of model parameters, about 99% of the total sample variation in shipping cost (y) is explained by the model; the remainder is explained by random error.

f) Is the model statistically useful for the prediction of shipping cost y? Find the value of the F statistic on the printout and give the observed significance level (p-value) for the test.

F-statistic: 458.4 on 5 and 14 DF, p-value: $5.371e - 15$

Since p-value is very small, we reject H_0 and that conclude that the model is useful for predicting shipping cost y.

g) Find a 95% prediction interval for the cost of shipping a 5-pound package a distance of 100 miles.



```
> New=data.frame(Weight=c(5), Distance=(100))
> predict(model,New)
       1
4.241419
> predict(model,New, interval="confidence",level=0.95)
       fit      lwr      upr
1 4.241419 3.847502 4.635336
> predict(model,New, interval="prediction",level=0.95)
       fit      lwr      upr
1 4.241419 3.213297 5.269541
```

Figure 4.6: 95% prediction interval and confidence interval

Interpretation:

Models with Qualitative x's

Categorical (qualitative) variables can also be added to our multiple regression model. Since these independent variables cannot be measured on a numerical scale, we need to code the categories (levels) of the categorical variables as numebrs.

Dummy Variables

Coded categorical variables are called dummy variables.

For example, x might represent sex where $x = 0$ indicates male and $x = 1$ indicates female.

Note: Using 0 and 1 to code the values of the categorical variable helps us to interprete β easier.

In this case, x is called a dummy or indicator variable.

$$x = \begin{cases} 0 & \text{if individual is a male} \\ 1 & \text{if individual is a female} \end{cases}$$

A Model Relating $E(y)$ to a Qualitative Independent Variable with Two Levels

$$E(y) = \beta_0 + \beta_1 x$$

where:

$$x = \begin{cases} 1 & \text{if level A} \\ 0 & \text{if level B} \end{cases}$$

Interpretation of β's:

$$\beta_0 = \mu_B \quad \text{(Mean for base level)}$$

$$\beta_1 = \mu_A - \mu_B$$

A Model Relating $E(y)$ to a Qualitative Independent Variable with Three Levels

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where:

$x_1 = \begin{cases} 1 & \text{if level A} \\ 0 & \text{if not} \end{cases}$ $\quad x_2 = \begin{cases} 1 & \text{if level B} \\ 0 & \text{if not} \end{cases}$ $\quad$ Base Level = Level C

Interpretation of β's:

$$\beta_0 = \mu_C \quad \text{(Mean for base level)}$$

$$\beta_1 = \mu_A - \mu_C$$

$$\beta_2 = \mu_B - \mu_C$$

Note:

- The number of dummy variables used to describe the categorical variable will be one less than the number of the levels of the categorical variable.
- **Base Level** is the level of the categorical variable assigned the value 0. β_0 always represents the mean response assosiated with the base level.

The number of dummy variables used to describe the categorical variable will be one less than the number of the levels of the categorical variable.

Dummy Variable Model for 1 Qualitative x

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{k-1} x_{k-1}$$

where:

$x_1 = \begin{cases} 1 & \text{if level 1} \\ 0 & \text{if not} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if level 2} \\ 0 & \text{if not} \end{cases} \quad ...x_{k-1} = \begin{cases} 1 & \text{if level k-1} \\ 0 & \text{if not} \end{cases}$

Interpretation of β's:

$$\beta_0 = E(y) \text{ for level k (base level)} = \mu_k$$

$$\beta_1 = \mu_1 - \mu_k$$

$$\beta_2 = \mu_2 - \mu_k$$

Example 4.12.2 *Refer to Exercise 4.53 from the textbook, "Homework assistance for accounting students." Data set: ACCHW*

A study about the best method of assisting accounting students with their homework. 75 accounting students took a pretest on a topic not covered in class, then each was given a home- work problem to solve on the same topic. The students were assigned to one of three home- work assistance groups. Some students received the completed solution, some were given check figures at various steps of the solution, and some received no help at all. After finishing the home- work, the students were all given a posttest on the subject. The dependent variable of interest was the knowledge gain (or test score improvement).

In this data set we have a numerical dependent variable and one categorical independent variable with 3 levels: No help, Check figures, Full solutions. Therefore, we need $3 - 1 = 2$ dummy variables. We arbitrarily select "No help" as the base level. Therefore, the other two levels are assigned 1 or 0 depending on the dummy variable.

$x_1 = \begin{cases} 1 & \text{if FULL} \\ 0 & \text{if not} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if CHECK} \\ 0 & \text{if not} \end{cases} \quad$ Base Level = NO

(a) Propose a model for the knowledge gain (y) as a function of the qualitative variable, homework assistance group.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where $\beta_0 = \mu_{NO}$ when $x_1 = x_2 = 0$

(b) In terms of the β's in the model, give an expression for the difference between the mean knowledge gains of students in the "completed solution" and "no help groups."

$x_2 = 0 \longrightarrow \mu_{FULL} = \beta_0 + \beta_1 \longrightarrow \mu_{FULL} = \mu_{NO} + \beta_1 \longrightarrow \beta_1 = \mu_{FULL} - \mu_{NO}$

(c) Interpret the estimated β coefficients in the model.

We need to write the mean Knowledge Gain, $E(y)$ for each of the three levels of home- work assistance as a function of the β's:

FULL: $x_1 = 1, x_2 = 0$

$E(y) = \beta_0 + \beta_1(1) + \beta_2(0) \longrightarrow \mu_{FULL} = \beta_0 + \beta_1$

CHECK: $x_1 = 0, x_2 = 1$

$E(y) = \beta_0 + \beta_1(0) + \beta_2(1) \longrightarrow \mu_{CHECK} = \beta_0 + \beta_2$

NO: $x_1 = 0, x_2 = 0$

$E(y) = \beta_0 + \beta_1(0) + \beta_2(0) \longrightarrow \mu_{NO} = \beta_0$

Then we have:

$\beta_0 = \mu_{NO}$ (Mean of the base level)

$\beta_1 = \mu_{FULL} - \mu_{NO}$

$\beta_2 = \mu_{CHECK} - \mu_{NO}$

According to the output:

$\hat{\beta_0} = 2.4333$, $\hat{\beta_1} = -0.4833$, and $\hat{\beta_2} = 0.2867$

Therefore,

The estimated mean knowledge gain (or test score improvement) for students who received no help is 2.4333 points.

The difference between the estimated mean knowledge gain (or test score improvement) for students who received the completed solution and for those who received no help is -0.4833 points.

The difference between the estimated mean knowledge gain (or test score improvement) for students who were given check figures at various steps of the solution and for those who received no help is 0.2867 points.

(d) Fit the model to the data and give the least squares prediction equation.

Predicted Knowledge Gain $= 2.4333 - 0.4833 \times x_1 + 0.2876 \times x_2$

(d) Conduct the global F -Test for model utility using $\alpha = .05$. Interpret the results, practically.

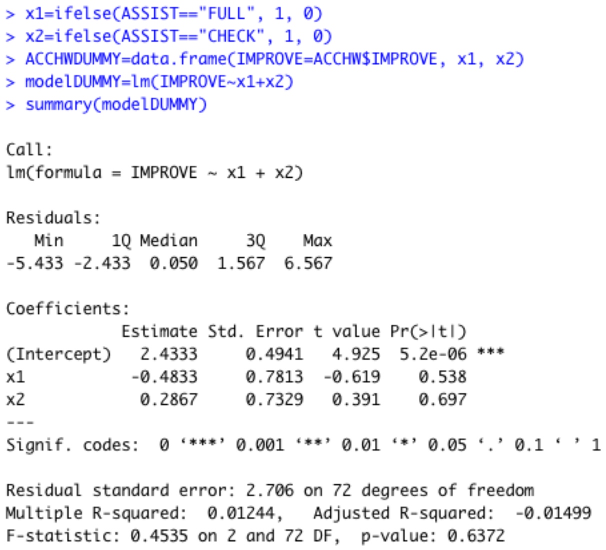
$H_0 = \beta_1 = \beta_2 = 0$

$H_a =$ at least one $\beta_i =\neq 0$

$F_c = 0.4535$ and p-value $= 0.6372 > 0.05 \longrightarrow$ we fail to reject H_0.

At 5% level of significance, we do not have sufficient evidence to conclude that the model is significant in predicting the knowledge gained.

Note: $\beta_1 = \beta_2 = 0$ implies that $\mu_{FULL} = \mu_{CHECK} = \mu_{NO}$



```
> x1=ifelse(ASSIST=="FULL", 1, 0)
> x2=ifelse(ASSIST=="CHECK", 1, 0)
> ACCHWDUMMY=data.frame(IMPROVE=ACCHW$IMPROVE, x1, x2)
> modelDUMMY=lm(IMPROVE~x1+x2)
> summary(modelDUMMY)

Call:
lm(formula = IMPROVE ~ x1 + x2)

Residuals:
   Min     1Q Median     3Q    Max 
-5.433 -2.433  0.050  1.567  6.567 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.4333     0.4941   4.925  5.2e-06 ***
x1           -0.4833     0.7813  -0.619    0.538    
x2            0.2867     0.7329   0.391    0.697    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.706 on 72 degrees of freedom
Multiple R-squared:  0.01244,	Adjusted R-squared:  -0.01499 
F-statistic: 0.4535 on 2 and 72 DF,  p-value: 0.6372
```

Figure 4.7: Multiple Regression Output

4.13 A Test for Comparing Nested Models

Suppose we would like to determine which model best fits the data with a high degree of confidence.

Nested Models

Two models are **nested** if one model contains all the terms of the second model and at least one additional term. The more complex of the two models is called the **complete (or full) model**. The simpler of the two models is called the **reduced (or restricted) model**.

Consider the following two possible models:

1. Straight-line interaction model (reduced model): $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

2. Curvilinear model (complete (or full) model.): $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$

Question: Does the curvilinear (or complete) model contributes more information for the prediction of y than the straight-line (or reduced) model?

To answer this question, we conduct the following test:

$H_0 : \beta_4 = \beta_5 = 0$

H_a : At least one of the parameters, β_4 or β_5 differs from 0

1- We use the method of least squares to fit the reduced model and calculate the corresponding sum of squares for error, SSE_R

2- We fit the complete model and calculate its sum of squares for error, SSE_C. Then, we compare SSE_R to SSE_C by calculating the difference, $SSE_R - SSE_C$.

The larger the difference, the greater the weight of evidence that the complete model provides better predictions of y than does the reduced model.

We use an F statistic:

$F = \frac{\frac{\text{Drop in SSE}}{\text{Number of parameters being tested}}}{s^2\text{(for larger model)}} = \frac{\frac{(SSE_R - SSE_C)}{2}}{\frac{SSE_c}{[n-(5+1)]}}$

$d.f. = (\nu_1, \nu_2)$ when $\nu_1 = 2$ and $\nu_2 = n - 6$

F-Test for Comparing Nested Models

Reduced model: $E(y) = \beta_0 + \beta_1 x_1 + ... + \beta_g x_1 x_g$

Complete (or full) model: $E(y) = \beta_0 + \beta_1 x_1 + ... + \beta_g x_g + \beta_{g+1} x_{g+1} + ... + \beta_k x_k$

$H_0 : \beta_{g+1} = \beta_{g+2} + ... + \beta_k = 0$

H_a : At least one of the parameters differs from 0

$$F_c = \frac{\frac{(SSE_R - SSE_C)}{k-g}}{\frac{SSE_c}{[n-(k+1)]}} = \frac{\frac{SSE_R - SSE_C}{\text{Number of parameters being tested}}}{MSE_C}$$

Where:

$SSE_R =$ Sum of squared errors for the reduced model

$SSE_C =$ Sum of squared errors for the complete model

$MSE_C =$ Mean square error for the complete model

$k - g =$ Number of β parameters specified in H_0

$k + 1 =$ Number of β parameters in the complete model (including β_0

$n =$ Total sample size

Rejection region Approach: $F > F_\alpha(\nu_1, \nu_2)$

where:

$\nu_1 = k - g$ is degrees of freedom for the numerator

$\nu_2 = n - (k + 1)$ is degrees of freedom for the denominator

p-value Approach:

p-value $= P(F > F_c)$ H_0 is rejected when p-value $< \alpha$

Example 4.13.1 *Refer to* ***Cooling method for gas turbines*** *example. Consider a model for heat rate (kilojoules per kilowatt per hour) of a gas turbine as a function of cycle speed (revolutions per minute) and cycle pressure ratio. The data are saved in the GASTURBINE file.*

a) Write a complete second-order model for heat rate (y).

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$

Variables:

y, HEATRATE: Heat rate (kilojoules per kilowatt per hour).

x_1, RPM: Cycle Speed: (revolutions per minute) of the engine.

x_2, CPRATIO: Cycle Pressure Ratio.

b) Give the null and alternative hypotheses for determining whether the curvature terms in the complete second-order model are statistically useful for predicting heat rate (y).

$H_0 : \beta_4 = \beta_5 = 0$

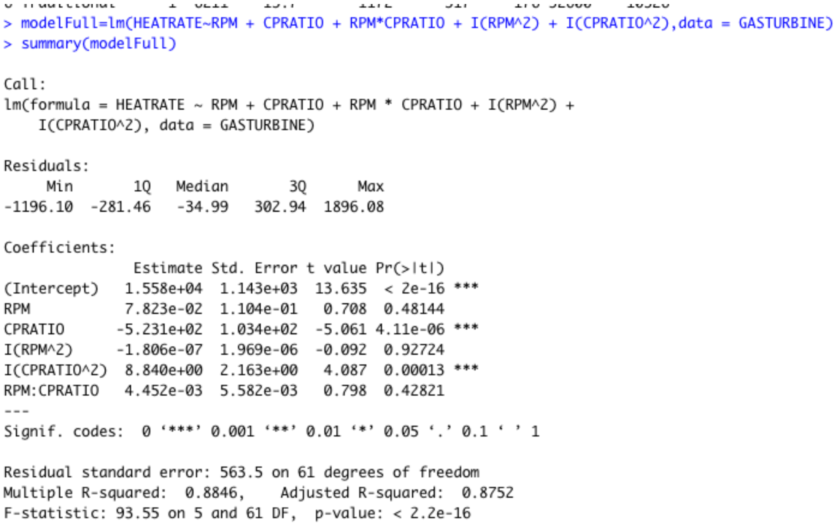
H_a : At least one of the parameters, β_4 or β_5 differs from 0

c) For the test in part b, identify the "complete" and "reduced" model.

1. Reduced model: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + +\beta_3 x_1 x_2$

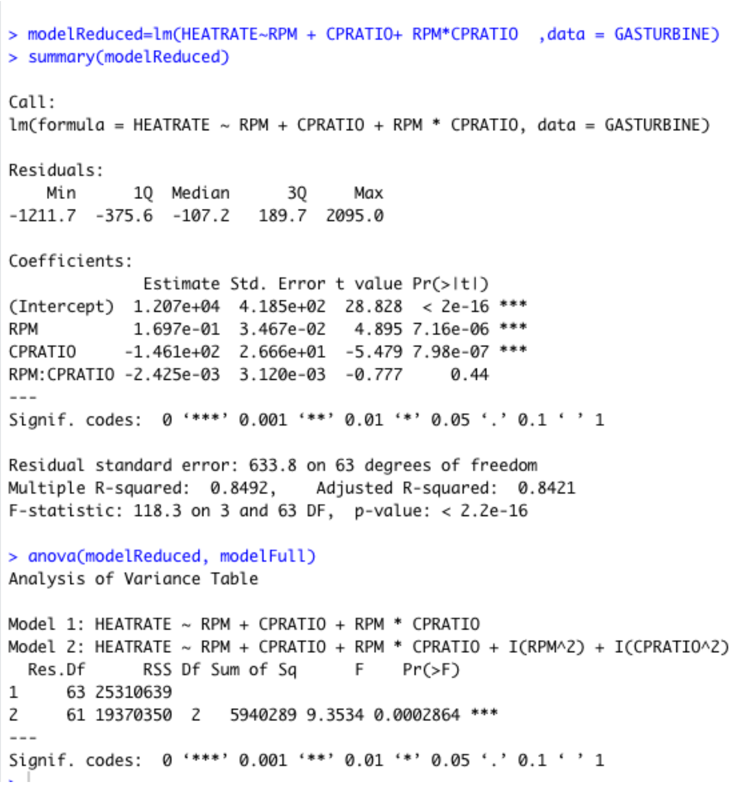
2. Curvilinear model (complete (or full) model): $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$

The following outputs are for the full, reduced models, and partial F test respectively:



```
> modelFull=lm(HEATRATE~RPM + CPRATIO + RPM*CPRATIO + I(RPM^2) + I(CPRATIO^2),data = GASTURBINE)
> summary(modelFull)

Call:
lm(formula = HEATRATE ~ RPM + CPRATIO + RPM * CPRATIO + I(RPM^2) +
    I(CPRATIO^2), data = GASTURBINE)

Residuals:
     Min       1Q   Median       3Q      Max
-1196.10  -281.46   -34.99   302.94  1896.08

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.558e+04  1.143e+03  13.635  < 2e-16 ***
RPM           7.823e-02  1.104e-01   0.708  0.48144
CPRATIO      -5.231e+02  1.034e+02  -5.061 4.11e-06 ***
I(RPM^2)     -1.806e-07  1.969e-06  -0.092  0.92724
I(CPRATIO^2)  8.840e+00  2.163e+00   4.087  0.00013 ***
RPM:CPRATIO   4.452e-03  5.582e-03   0.798  0.42821
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 563.5 on 61 degrees of freedom
Multiple R-squared:  0.8846,    Adjusted R-squared:  0.8752
F-statistic: 93.55 on 5 and 61 DF,  p-value: < 2.2e-16
```

Figure 4.8: Full-Multiple Regression Output



```
> modelReduced=lm(HEATRATE~RPM + CPRATIO+ RPM*CPRATIO  ,data = GASTURBINE)
> summary(modelReduced)

Call:
lm(formula = HEATRATE ~ RPM + CPRATIO + RPM * CPRATIO, data = GASTURBINE)

Residuals:
    Min      1Q  Median      3Q     Max
-1211.7  -375.6  -107.2   189.7  2095.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.207e+04  4.185e+02  28.828  < 2e-16 ***
RPM          1.697e-01  3.467e-02   4.895 7.16e-06 ***
CPRATIO     -1.461e+02  2.666e+01  -5.479 7.98e-07 ***
RPM:CPRATIO -2.425e-03  3.120e-03  -0.777     0.44
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 633.8 on 63 degrees of freedom
Multiple R-squared:  0.8492,    Adjusted R-squared:  0.8421
F-statistic: 118.3 on 3 and 63 DF,  p-value: < 2.2e-16

> anova(modelReduced, modelFull)
Analysis of Variance Table

Model 1: HEATRATE ~ RPM + CPRATIO + RPM * CPRATIO
Model 2: HEATRATE ~ RPM + CPRATIO + RPM * CPRATIO + I(RPM^2) + I(CPRATIO^2)
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1     63 25310639
2     61 19370350  2   5940289 9.3534 0.0002864 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

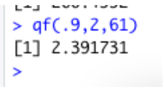
d) Find the values of SSE_R , SSE_C , and MSE_C.

$SSE_R = 25310639$, $SSE_C = 19370350$, and $MSE_C = \frac{19370350}{61} = 317546.721$

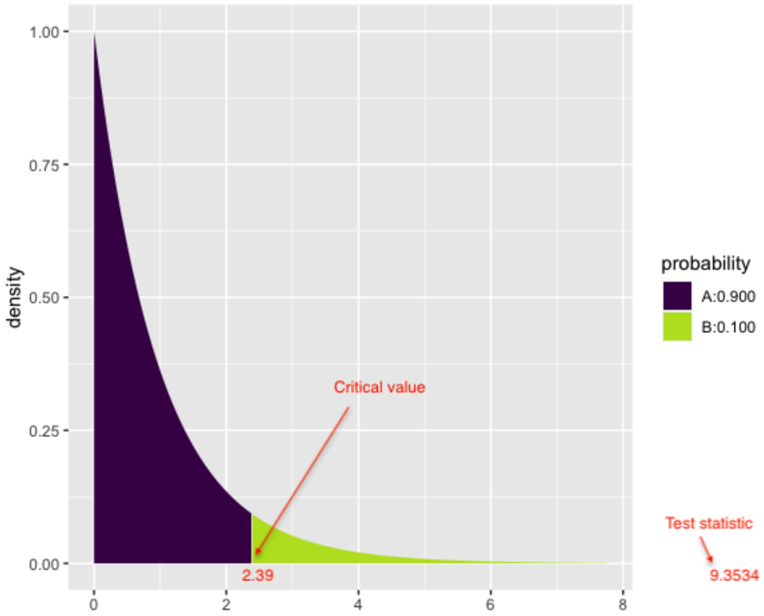
e) Compute the value of the test statistics for the test of part (b).

$$F_c = \frac{\frac{SSE_R - SSE_C}{\text{Number of parameters being tested}}}{MSE_C} = \frac{\frac{25310639-19370350}{2}}{317546.721} = 9.353408$$

f) Find the rejection region for the test of part (b) using $\alpha = .10$.



```
> qf(.9,2,61)
[1] 2.391731
>
```



g) State the conclusion of the test in the words of the problem.

Since the test statistic, $F = 9.353408$ falls in rejection region, we reject H_0.

At 10% level of significance, we have sufficient evidence that the quadratic terms contribute to the prediction of HEATRate.

Note:

- We must be cautious about failing to reject H_0 since the probability of a Type II error is unknown.
- If two competing models are found to have essentially the same predictive power, the model with the **lower** number of β's (i.e., the more **parsimonious** model) is selected.
- If the models are not nested, this F -Test is not applicable. In this situation, the analyst must base the choice of the best model on statistics such as R_a^2 and s.

A **parsimonious model** is a model with a small number of β parameters. In situations where two competing models have essentially the same predictive power (as determined by an F -Test), choose the more parsimonious of the two.

4.14 A Complete Example

In US, commercial contractors bid for the right to construct state highways and roads. A state government agency, usually the Department of Trans- portation (DOT), notifies various contractors of the state's intent to build a highway. Contractor with the lowest bid (building cost) is awarded the road construction contract. Our objective is to build and test the adequacy of a model designed to predict the cost y of a road construction contract awarded using the sealed-bid system in Florida.

Data set: FLAG, $n = 235$

Dependent variable, y: Contract Cost (in thousands of dollars)

Independent variables:

1. x_1: The DOT engineer's estimate of the cost (in thousands of dollars), Numerical
2. x_2: Status of the bid contract, (Fixed, Competitive), Categorical

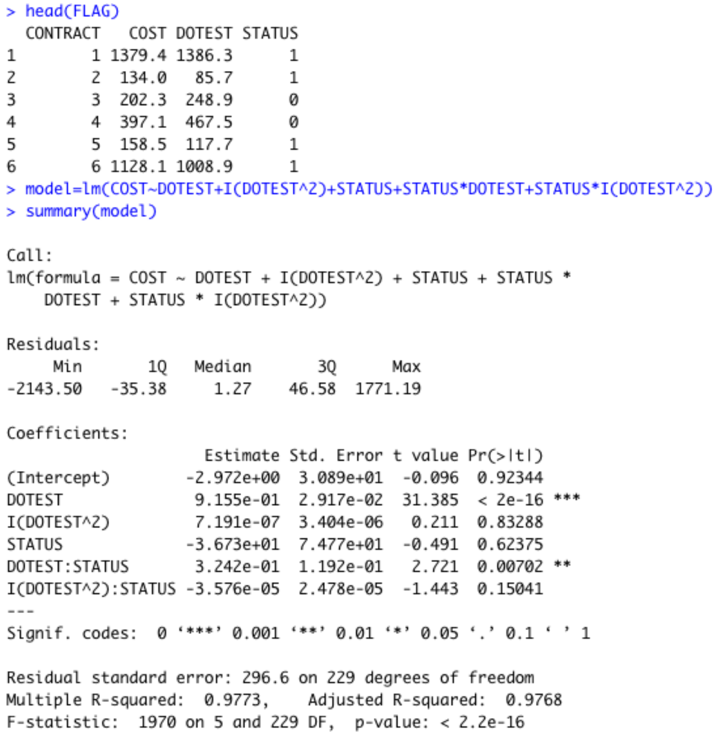
Status is a categorical variable with two levels, so we need one dummy variable,

$$x_2 = \begin{cases} 1 & \text{if Fixed} \\ 0 & \text{if not} \end{cases}$$

Base Level = Competitive

A complete second- order model with one two levels categorical and one numerical variable

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_1 x_2 + \beta_5 x_1^2 x_2$$



```
> head(FLAG)
  CONTRACT   COST DOTEST STATUS
1        1 1379.4 1386.3      1
2        2  134.0   85.7      1
3        3  202.3  248.9      0
4        4  397.1  467.5      0
5        5  158.5  117.7      1
6        6 1128.1 1008.9      1
> model=lm(COST~DOTEST+I(DOTEST^2)+STATUS+STATUS*DOTEST+STATUS*I(DOTEST^2))
> summary(model)

Call:
lm(formula = COST ~ DOTEST + I(DOTEST^2) + STATUS + STATUS *
    DOTEST + STATUS * I(DOTEST^2))

Residuals:
     Min       1Q   Median       3Q      Max
-2143.50   -35.38     1.27    46.58  1771.19

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -2.972e+00  3.089e+01  -0.096  0.92344
DOTEST              9.155e-01  2.917e-02  31.385  < 2e-16 ***
I(DOTEST^2)         7.191e-07  3.404e-06   0.211  0.83288
STATUS             -3.673e+01  7.477e+01  -0.491  0.62375
DOTEST:STATUS       3.242e-01  1.192e-01   2.721  0.00702 **
I(DOTEST^2):STATUS -3.576e-05  2.478e-05  -1.443  0.15041
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 296.6 on 229 degrees of freedom
Multiple R-squared:  0.9773,	Adjusted R-squared:  0.9768
F-statistic:  1970 on 5 and 229 DF,  p-value: < 2.2e-16
```

Figure 4.9: Complete Second- Order Mode for DATA FLAGl

According to the RStudio output above:

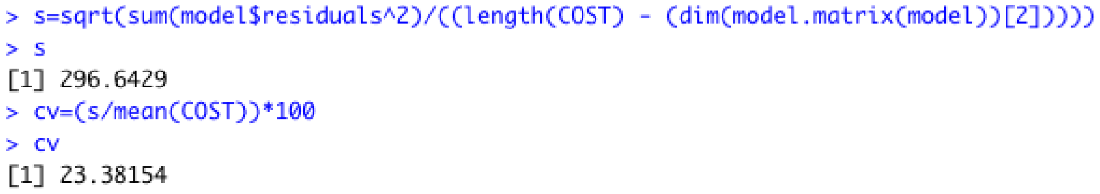
$\hat{\beta}_0 = -2.972$, $\hat{\beta}_1 = 0.9155$, $\hat{\beta}_2 = 0.0000007191$, $\hat{\beta}_3 = -36.73$, $\hat{\beta}_4 = 0.3242$, and $\hat{\beta}_5 = -0.00003576$

Interpretation of the model standard deviation, s

According to the RStudio output above:

Residual standard error, $s = 296.6$ therefore, about 95% of the contract costs for the sample of $n = 235$ contracts fall within $\$593,200$ $(2s)$ of the model predicted values.

To determine whether this is a reasonable potential error of prediction, we calculate $C.V.$ of which a value of %10 or smaller usually leads to more precise predictions.



```
> s=sqrt(sum(model$residuals^2)/((length(COST) - (dim(model.matrix(model))[2]))))
> s
[1] 296.6429
> cv=(s/mean(COST))*100
> cv
[1] 23.38154
```

The relatively high value of %23 of $C.V.$ leads to less precise predictions by having wide prediction intervals.

Although the extremely small p-value $< 2.2e-16$ indicates that the model ($H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ $vs.$ H_a : at least one $\beta_j \neq 0$) is statistically adequate (at $\alpha = .01$) for predicting contract cost, y, we are not sure whether all the terms in the model statistically significant predictors.

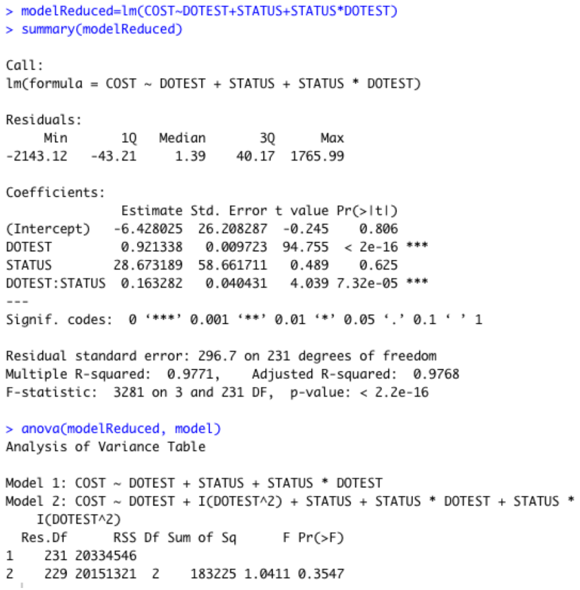
To determine whether the curvilinear terms should be included in the model or not, we test:

$H_0 : \beta_2 = \beta_5 = 0$

H_a : At least one of the curvature β's is nonzero.

Reduced Model:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_3 x_2 + \beta_4 x_1 x_2$$



```
> modelReduced=lm(COST~DOTEST+STATUS+STATUS*DOTEST)
> summary(modelReduced)

Call:
lm(formula = COST ~ DOTEST + STATUS + STATUS * DOTEST)

Residuals:
     Min       1Q   Median       3Q      Max 
-2143.12   -43.21     1.39    40.17  1765.99 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -6.428025  26.208287  -0.245    0.806    
DOTEST         0.921338   0.009723  94.755  < 2e-16 ***
STATUS        28.673189  58.661711   0.489    0.625    
DOTEST:STATUS  0.163282   0.040431   4.039 7.32e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 296.7 on 231 degrees of freedom
Multiple R-squared:  0.9771,    Adjusted R-squared:  0.9768 
F-statistic:  3281 on 3 and 231 DF,  p-value: < 2.2e-16

> anova(modelReduced, model)
Analysis of Variance Table

Model 1: COST ~ DOTEST + STATUS + STATUS * DOTEST
Model 2: COST ~ DOTEST + I(DOTEST^2) + STATUS + STATUS * DOTEST + STATUS * 
    I(DOTEST^2)
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1    231 20334546                           
2    229 20151321  2    183225 1.0411 0.3547
```

Figure 4.10: Reduced Model and partial F test

Since the p-value $= 0.3547 > 0.01$, we fail to reject H_0.

At %1 level of significance, there is insufficient evidence to indicate that the curvature terms are useful predictors of construction cost, y.

Therefore, the model is:

The predicted cost of contract $= -6.428025 + 0.921338x_1 + 28.673189x_2 + 0.163282x_1x_2$

Question: Can the model be simplified any further?

The following scatterplot depicts the relationship between cost and DOT estimate by taking into account the status of the bid contract, fixed or competitive.

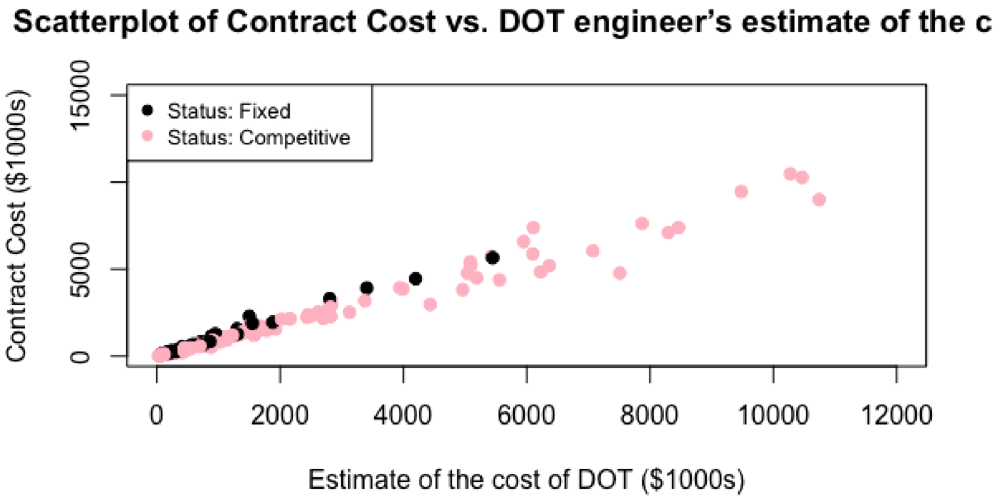


Figure 4.11: Scatterplot of Cost vs. Dot Estimate by Status

Now, let's fit the least squares lines for both fixed and competitive contracts:

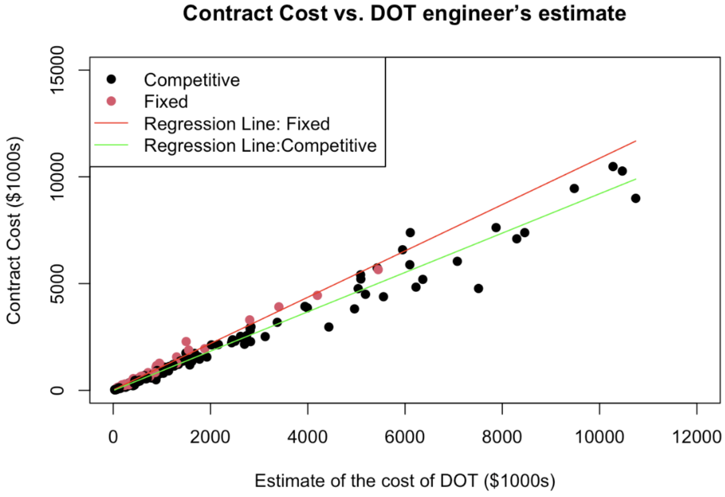


Figure 4.12: Plot of the least squares lines for the reduced model

The least squares lines are not parallel implying we have interaction.

Let's find the least squares lines for both fixed and competitive contracts:

The predicted cost of contract $= -6.428025 + 0.921338x_1 + 28.673189x_2 + 0.163282x_1x_2$

Fixed ($x_2 = 1$): $\hat{y} = -6.428025 + 0.921338x_1 + 28.673189(1) + 0.163282x_1(1) = 22.24516 + 1.08462x_1$

Competitive ($x_2 = 0$): $\hat{y} = -6.428025 + 0.921338x_1 + 28.673189(0) + 0.163282x_1(0) = -6.42802 + 0.92134x_1$

Interpretations:

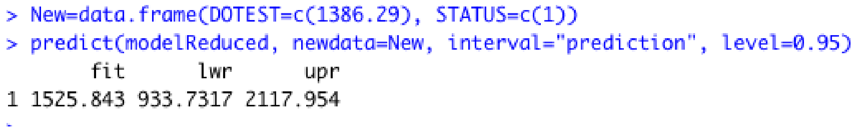
Fixed: On average, for every \$1000 increase in the DOT engineer's estimate of the cost, the cost of contract will be estimated to increase by \$1084.62 for fixed contracts.

Competitive: On average, for every \$1000 increase in the DOT engineer's estimate of the cost, the cost of contract will be estimated to increase by \$921.338 for competitive contracts.

Interpretation of R_a^2

The value of $R_a^2 = 0.9768$ means about %98 of the variation in the sample of construction costs can be "explained" by the model. Although the value of R_a^2 is large, the value of $s = 296.6957$ implies that we can predict construction cost to within about $2s = 2(296.6957) = 593.3914$ thousand dollars of its true value using the model. Therefore, the predictive ability of the model might be improved by additional independent variables.

The following output shows the %95 prediction interval for contract cost when the DOT estimateis $\$1,386,290 (x_1 = 1,386.29)$ and the contract is fixed ($x_2 = 1$).



```
> New=data.frame(DOTEST=c(1386.29), STATUS=c(1))
> predict(modelReduced, newdata=New, interval="prediction", level=0.95)
       fit      lwr      upr
1 1525.843 933.7317 2117.954
```

According to the interval above, with 95% confidence, we predict that the cost of fixed contract falls between $\$933,731.7$ and $\$2,117,954$ when the DOT estimateis $\$1,386,290$.

Note:

1. If curvature (x^2) deemed important, do not conduct test for first-order (x) term in the model.

2. If interaction (x_1x_2) deemed important, do not conduct tests for first-order terms (x_1 and x_2) in the model.

Recommendation for Assessing Model

1. Conduct global F -test; if significant then:

2. Conduct t -tests on only the most important β's (interaction or squared terms)

3. Interpret value of $2s$ considering $C.V.$

4. Interpret value of R_a^2

Acknowledgement

The core content of the slides are from the textbook of this course;

A Second Course in Statistics: Regression Analysis (7th Edition)

by

Mendenhall, William and Sincich, Terry; Pearson Education.