Contents

Chapter 7

Some Regression Pitfalls

In this chapter, we identify some potential problems while constructing a model for a response variable, y and avoid some of the problems by recognizing them.

7.1 Introduction

Multiple regression is a powerful tool when we incorporate other potentially important independent variables into the model to make accurate predictions. However, there are a number of pitfalls that we should be aware of when constructing a multiple regression model.

7.2 Observational Data versus Designed Experiments

Recall that the data for regression can be either observational (where the values of the independent variables are uncontrolled) or experimental (where the x's are controlled via a designed experiment).

1- In an experiment, the quantity of information is controlled not only by the amount of data, but also by the values of the predictor variables x_1 , x_2 , . . . , x_k. Hence, we might be able to increase the amount of information in the data by designing the experiment at no additional cost.

2- When we use observational data, a problem involving randomization creates.

In designed experiments, the experimental units are randomly selected for each combination of the independent variables. This procedure tends to average out any variation within the experimental units. If the difference between two sample means is statistically significant, then you can infer (considering Type I error) that the population means differ and more important you can infer a cause-and-effect relationship wich cannot be infered when you have observational data and simply means that x contributes information for the prediction of y.

Note: With observational data, a statistically significant relationship between a response y and a predictor variable x does not necessarily imply a cause-and- effect relationship.

7.3 Parameter Estimability and Interpretation

Suppose we want to fit the first-order model:

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

When:

$y =$ A developmentally challenged child's creativity score

$x =$ The child's flexibility score

$n = 3$

Suppose $x_i = 5, 5, 5$. The problem is that the parameters of the straight-line model cannot be estimated when all the data are concentrated at a single x-value.

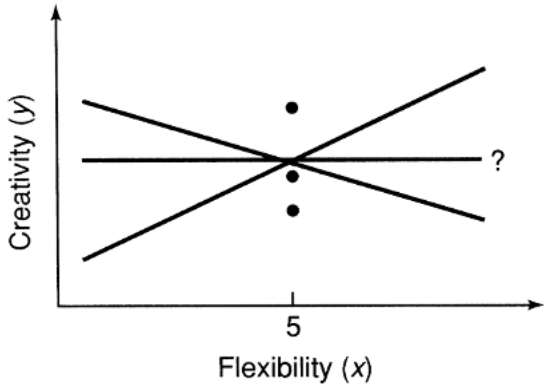


Figure 7.1: Creativity and flexibility data for three children

A similar problem would occur if we attempted to fit the second-order model to a set of data for which only one or two different x-values were observed.

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

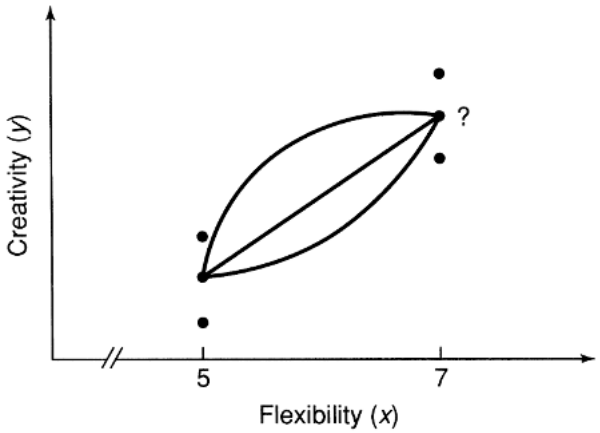


Figure 7.2: Only two different x-values observed - the second-order model is not estimable

Requirements for Fitting a pth-Order Polynomial Regression Model

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

1. The number of levels of x must be greater than or equal to $(p + 1)$.
2. The sample size n must be greater than $(p + 1)$ to allow sufficient degrees of freedom for estimating σ^2.

7.4 Multicollinearity

When two or more of the independent variables in the model for $E(y)$ are correlated with each other, they will contribute redundant information and we say that **multicollinearity** exists.

Multicollinearity exists when two or more of the independent variables used in regression are moderately or highly correlated.

Problems with extreme multicollinearity

1. Increase in the likelihood of rounding errors in the calculations of the β estimates, standard errors, and so forth.
2. The regression results may be confusing and misleading.
3. Effect on the signs of the parameter estimates.

Ways to avoid multicollinearity

- Conducting a designed experiment.
- Calculating the coefficient of correlation r between each pair of independent variables in the model.
- Nonsignificant t-tests for the individual β parameters when the F-test for overall model adequacy is significant.
- Estimating with opposite signs than expected for the parameters.

VIF

Variance Inflation Factors is a method for detecting multicollinearity that calculates variance inflation factors for the individual β parameters.

Detecting Multicollinearity in the Regression Model

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

The following are indicators of multicollinearity:

1. Significant correlations between pairs of independent variables in the model
2. Nonsignificant t-tests for all (or nearly all) the individual β parameters when the F-test for overall model adequacy $H_0 : \beta_1 = \beta_2 = \Delta\Delta\Delta = \beta_k = 0$ is significant
3. Opposite signs (from what is expected) in the estimated parameters
4. A variance inflation factor (VIF) for a β parameter greater than 10, where

$$VIF = \frac{1}{1 - R_i^2}; i = 1, 2, ..., k$$

R_i^2 is the multiple coefficient of determination for the model

$E(x_i) = \alpha_0 + \alpha_1 x_1 + ... + \alpha_{i-1} x_{i-1} + \alpha_{i+1} x_{i+1} + ... + \alpha_k x_k$

Example 7.4.1 *Refer to Example 7.5 from the text book. Data set: FTCCIGAR*

$n = 25$, $y =$ Carbon Monoxide Content (in milligrams)

$x_1 =$ tar content (in milligrams)

$x_2 =$ nicotine content (in milligrams)

$x_3 =$ weight (in grams)

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

First, we use the scatter-plot-matrix to plot the sample data and fit the model to the data set.

Consider the following $F-$test:

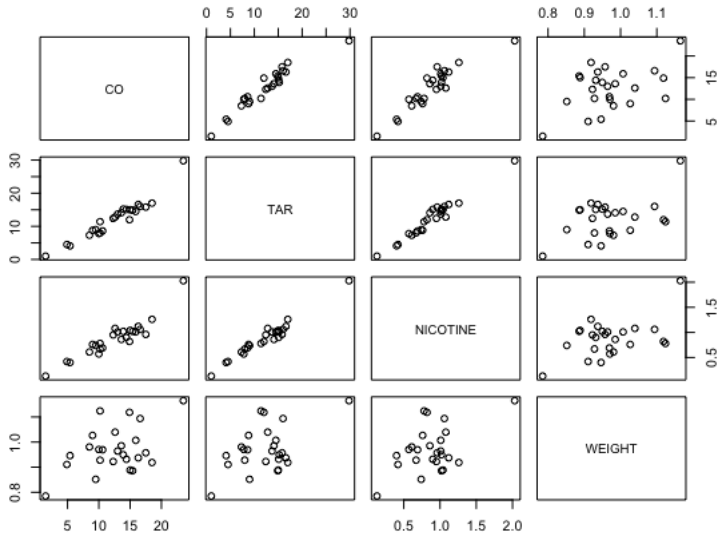
$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

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

The $F-$ statistic= 78.98 and $p-$value $= 1.329e-11 < \alpha = 0.05 \longrightarrow$ we reject H_0.

Conclusion: At 5% level of significance, The date provide strong evidence to conclude that the model is useful in predicting the carbon monoxide content.

The result of the F-test does not agree with the t-tests unless tar is the only one of the three variables useful for predicting carbon monoxide content.




```
> model1=lm(CO~TAR+NICOTINE+WEIGHT,data = FTCCIGAR)
> summary(model1)

Call:
lm(formula = CO ~ TAR + NICOTINE + WEIGHT, data = FTCCIGAR)

Residuals:
     Min       1Q   Median       3Q      Max
-2.89261 -0.78269  0.00428  0.92891  2.45082

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2022     3.4618   0.925 0.365464
TAR           0.9626     0.2422   3.974 0.000692 ***
NICOTINE     -2.6317     3.9006  -0.675 0.507234
WEIGHT       -0.1305     3.8853  -0.034 0.973527
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.446 on 21 degrees of freedom
Multiple R-squared:  0.9186,    Adjusted R-squared:  0.907
F-statistic: 78.98 on 3 and 21 DF,  p-value: 1.329e-11
```

The second clue is the negative value for $\hat{\beta}_2$ and $\hat{\beta}_3$ which is not the same as what we described from the scatterplot.

The following output shows the variance inflation factors for each of the three parameters.

```
> vif(model1)
      TAR  NICOTINE    WEIGHT
21.630706 21.899917  1.333859
>
```

Note that the variance inflation factors for both the tar and nicotine parameters are greater than 10.

The variance inflation factor for the tar parameter, $(VIF)_1 = 21.63$, implies that a model relating tar content x_1 to the remaining two independent variables, nicotine content x_2 and weight x_3, resulted in a coefficient of determination

$R_1 = 1 - \frac{1}{(VIF)_1} = 1 - \frac{1}{21.63} = .954$

The following output shows the coefficient of correlation r for each of the three pairs of independent variables in the model. All three sample correlations are significantly different from 0 based on the small p-values.

```
> X=cbind(TAR,NICOTINE,WEIGHT)
> rcorr(X, type="pearson")
          TAR NICOTINE WEIGHT
TAR      1.00     0.98   0.49
NICOTINE 0.98     1.00   0.50
WEIGHT   0.49     0.50   1.00

n= 25


P
         TAR    NICOTINE WEIGHT
TAR             0.0000   0.0127
NICOTINE 0.0000          0.0109
WEIGHT   0.0127 0.0109
```

Note: If the model will be used only for estimation and prediction, you may decide not to drop any of the independent variables from the model.

Solutions to Some Problems Created by Multicollinearity

1. Drop one or more of the correlated independent variables from the final model. A screening procedure such as stepwise regression is helpful in determining which variables to drop.
2. If you decide to keep all the independent variables in the model:
   a. Avoid making inferences about the individual β parameters (such as establishing a cause-and-effect relationship between y and the predictor variables).
   b. Restrict inferences about $E(y)$ and future y-values to values of the independent variables that fall within the experimental region (see Section 7.5).
3. If your ultimate objective is to establish a cause-and-effect relationship between y and the predictor variables, use a designed experiment (see Chapters 11 and 12).
4. To reduce rounding errors in polynomial regression models, code the independent variables so that first-, second-, and higher-order terms for a particular x variable are not highly correlated (see Section 5.6).
5. To reduce rounding errors and stabilize the regression coefficients, use ridge regression to estimate the β parameters (see Section 9.7).

7.5 Extrapolation: Predicting Outside the Experimental Region

The following figure shows the model may be very accurate for predicting y when x is in the range of experimentation, but the use of the model outside that range is a dangerous (although sometimes unavoidable) practice.

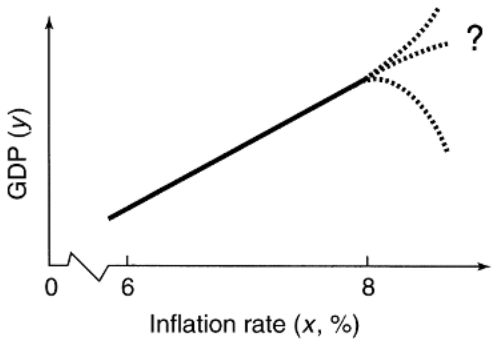


Figure 7.3: Using a regression model outside the experimental region

A $100(1-\alpha)\%$ prediction interval for GDP when the inflation rate is, say, 10%, will be less reliable than the stated confidence coefficient $(1-\alpha)$.

Establishing the experimental region for a multiple regression model that includes a number of independent variables may be more difficult.

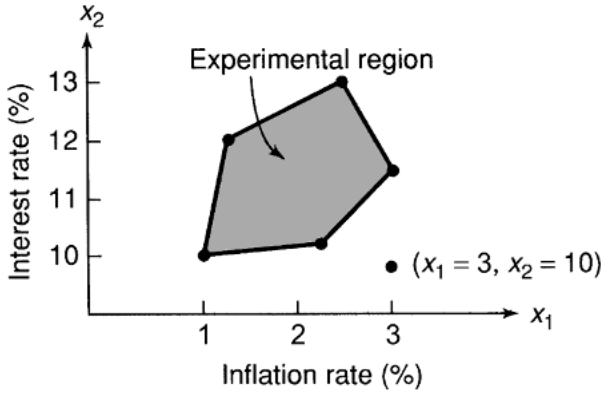


Figure 7.4: Experimental region for modeling GDP (y) as a function of inflation rate (x_1) and prime interest rate (x_2)

$n = 5$

x_1 ranges 1% to 3% and x_2 ranges 10% to 13% in the sample data.

the levels of x_1 and x_2 jointly define the region.

Note: Using the model to predict GDP for this observation — called **hidden extrapolation** — may lead to unreliable results.

7.6 Variable Transformations

Suppose x is one of the independent variables in a model is the price p of a commodity, we might choose to introduce this variable into the model as $x = \frac{1}{p}$ or $x = \sqrt{p}$ or $x = e^{-p}$. Thus, if we were to let $x = \sqrt{p}$, we would compute the square root of each price value, and these square roots would be the values of x that would be used in the regression analysis.

Transformations are performed on the y-values to make them more nearly satisfy the assumptions of the model (i.e. assumptions about random error) and, sometimes, to make the deterministic portion of the model a better approximation to the mean value of the transformed response.

Transformations of the values of the independent variables are performed solely to achieve a model that provides a better approximation to $E(y)$.

Consider the following graph that shows the relationship between the demand y and the price of a nonessential item. Suppose you expect the mean demand decreases as price p increases and then to decrease more slowly as p gets larger.

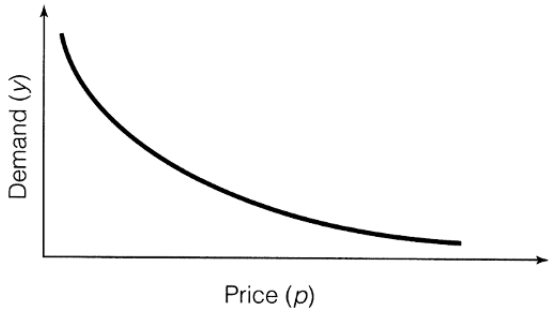


Figure 7.5: Hypothetical relation between demand y and price p

What function of p will provide a good approximation to $E(y)$?

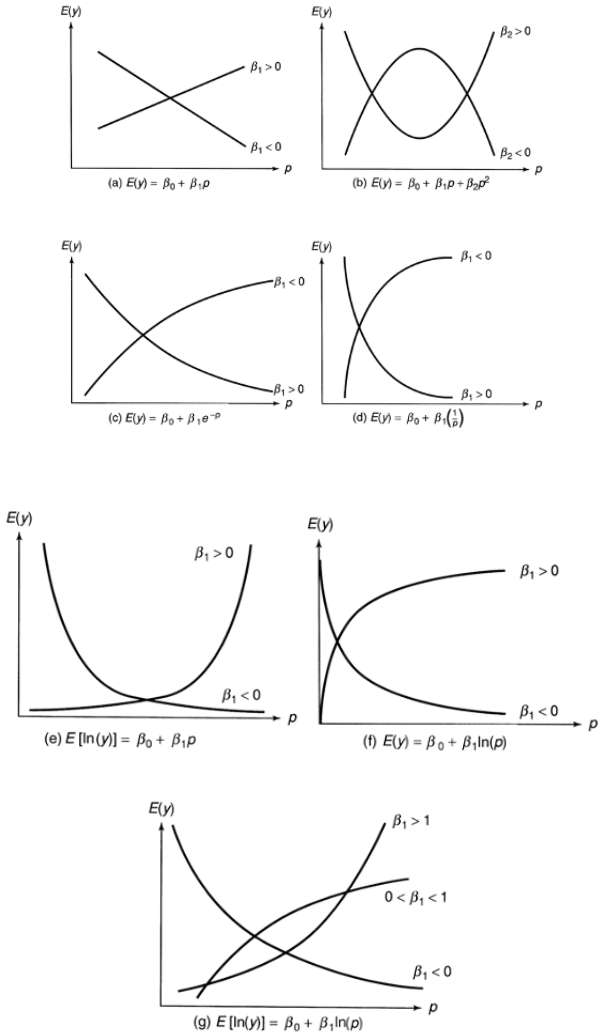


Figure 7.6: Graphs of some mathematical functions relating $E(y)$ to p

Example 7.6.1 *Refer to Example 7.8 from the text book. Data set: COFFEE*

$y =$ weekly demand (in pounds)

$p =$ price (in dollars)

$n = 8$

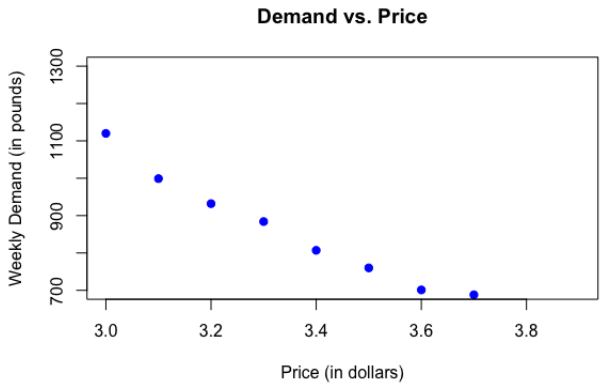


Figure 7.7: Scatterplot of weekly demand vs. price

a) Previous research by the supermarket chain indicates that weekly demand (y) decreases with price (p), but at a decreasing rate. This implies that model (d), Figure 7.11, is appropriate for predicting demand. Fit the model

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

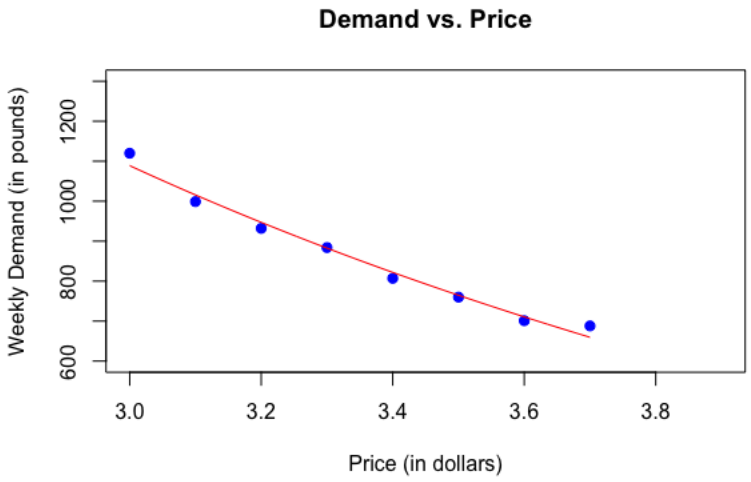
to the data, letting $x = \frac{1}{p}$.

```
> X=1/PRICE
> model2=lm(DEMAND~X)
> summary(model2)

Call:
lm(formula = DEMAND ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-16.681 -14.940  -7.175   8.178  31.113

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1180.5      107.7  -10.96 3.43e-05 ***
X             6808.1      358.4   19.00 1.37e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.9 on 6 degrees of freedom
Multiple R-squared:  0.9836,    Adjusted R-squared:  0.9809
F-statistic: 360.9 on 1 and 6 DF,  p-value: 1.375e-06
```

Figure 7.8: RStudio regression printout

$\hat{Demand} = -1180.5 + 6808.1X$

The following graph is the graph of this prediction equation.



(b) Do the data provide sufficient evidence to indicate that the model contributes information for the prediction of demand?

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

The $t = 19.00$ with the $p-$value $= 1.37e-06 < \alpha = 0.05 \longrightarrow$ we reject H_0 and and conclude that $x = \frac{1}{p}$ contributes information for the prediction of demand y.

(c) Find a 95% confidence interval for the mean demand when the price is set at \$3.20 per pound. Interpret this interval.

```
> New=data.frame(X=c(1/3.2))
> predict(model2, newdata=New, interval="confidence", level=0.95)
       fit      lwr      upr
1 947.0514 925.8664 968.2364
```

The interval is $(925.8664, 968.2364)$. Thus, we are 95% confident that mean demand will fall between 925.9 and 968.2 pounds when the price is set at \$3.20.

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.

A Modern Approach to Regression with R

by

Simon J. Sheather