Multiple Linear Regression & Other Considerations

Sections 3.2 & 6.1

Cathy Poliak, Ph.D.
cpoliak@central.uh.edu

Department of Mathematics
University of Houston

Outline

Recall The Example

The goal is to predict the *stock_index_price* (the dependent variable) of a fictitious economy based on three independent/input variables:

- *Interest_Rate*
- *Unemployment_Rate*
- *Year*

The data is in the *stock_price.csv* data set in BlackBoard. This is from `https://datatofish.com/multiple-linear-regression-in-r/`

We have looked at using interest rate as a predictor for the stock index price, what if we also add unemployment rate and year as predictors?

Can We Do Separate Simple Linear Regression Models?

Suppose now we also want to also include unemployement_rate as an input (predictor). Should we have two separate simple linear regression models?

- The approach of fitting a separate simple linear regression model for each predictor is not entirely satisfactory.
- It is unclear how to make a single prediction based on several models.
- Each of the separate models ignores the other predictors in forming estimates for the regression coefficients.
- Instead we extend the simple linear regression model so that it can directly accommodate multiple predictors.
- We give each predictor a separate slope coefficient in a single model.

General Form for Multiple Linear Regression

- Suppose we have p distinct predictors, the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p x_p + \epsilon$$

- X_j represents the jth predictor
- β_j quantifies the association between the jth predictor and the response.
- We interpret β_j as the **average** effect on Y of a one unit increase in X_j, **holding all other predictors fixed.**
- In our example of stock index price we have a model:

$$\text{stock\_index\_price} = \beta_0 + \beta_1 \times \text{Interest\_Rate} + \beta_2 \times \text{Unemployment\_Rate} + \beta_3 \times \text{Year} + \epsilon$$

Estimating the Regression Coefficients

- We now have p explanatory variables, we use the least-squares idea to find a linear function

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

- We use a subscript i to distinguish different cases. for the ith case the predicted response is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_p x_{ip}$$

- Using the *least squares method* we want $\hat{\beta}_j$ for $j = 1, \ldots, p$ that minimize

$$\begin{aligned} SSE &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

Linear Model of The Stock Index Price

```
stock3.lm <- lm(Stock_Index_Price~Interest_Rate+Unemployment_Rate+Year,
                data = stock_price)
summary(stock3.lm)

Call:
lm(formula = Stock_Index_Price ~ Interest_Rate + Unemployment_Rate +
Year, data = stock_price)

Residuals:
Min        1Q    Median       3Q      Max
-156.593  -41.552   -5.815   50.254  118.555

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -56523.71  134080.46  -0.422    0.678
Interest_Rate        324.59     123.37   2.631    0.016 *
Unemployment_Rate   -231.48     127.72  -1.812    0.085 .
Year                  28.89      66.42   0.435    0.668
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 71.96 on 20 degrees of freedom
Multiple R-squared:  0.8986,Adjusted R-squared:  0.8834
F-statistic: 59.07 on 3 and 20 DF,  p-value: 4.054e-10
```

$\hat{\beta}_0$ = $\hat{\beta}_1$ =

$H_0: \beta_j = 0$

$H_A: \beta_j \neq 0$

If p-value $\leq \alpha$

RH$_0$

$\therefore \beta_j X_j$ is a term in the model.

$$\widehat{stock\_index\_price} = -56523.71 + 324.59 \times Interest\_Rate - 231.48 \times Unemployement\_Rate + 28.89 \times Year$$

Interpretation of the Parameters

We interpret β_j as the average effect of Y (the predictor) of a one unit increase in X_j, **holding all other predictors fixed.**

- $\hat{\beta}_1 = 324.59$ This means that for 1% increase in interest rate, the stock index price will increase on average by \$324.48 for a fixed value of the unemployment rate and the year.

- $\hat{\beta}_2 = -231.48$, So for one 1% increase in unemployment rate, the stock index price will decrease on average by \$231.48 for a fixed value of the interest rate and the year.

- Give the interpretation of $\hat{\beta}_3$. $= 28.89$

For one year increase, the stock price will on average increase by \$28.89 for fixed interest rate and unemployment rate.

Correlation Matrix

```
> cor(stock_price[,-2])
                         Year Interest_Rate Unemployment_Rate Stock_Index_Price
Year                1.0000000     0.8828507        -0.8769997         0.8632321
Interest_Rate       0.8828507     1.0000000        -0.9258137         0.9357932
Unemployment_Rate  -0.8769997    -0.9258137         1.0000000        -0.9223376
Stock_Index_Price   0.8632321     0.9357932        -0.9223376         1.0000000
```

cor(Year, Interest rate) = 0.8828

cor(Year, unemployment rate) = -0.877

Be careful of multicolinarity!

- The occurance of high inter correlations among two or more independent variables.

Variance Inflation factor: VIF

Some Important Questions

For the **multivariate regression** we are interested in answering a few important questions.

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?

2. Do all of the predictors help to explain Y, or is only a subset of the predictors useful?

3. How well does the model fit the data?

4. Given a set of predictor values, what response value should we predict, and how acculturate is our prediction?

Answering the Questions

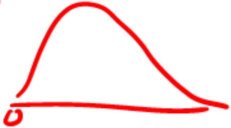
1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response? **Answer**: F - test, if p-value $\leq \alpha$ then at least one of the predictors are useful in predicting the response.

$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_p = 0$ $\quad y = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$

H_A: At least one $\beta_i \neq 0$ $\quad i = 1, 2, \ldots, p$

Test statistic: $F = \dfrac{SSR/p}{SSE/(n-p-1)}$

p = # of predictors
n = # of observations



p-value = $P(\text{Test Stat} \geq F_{p,\, n-p-1})$
Reject Ho if p-value is small

Answering the Questions

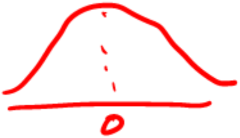
1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response? **Answer**: F - test, if p-value $\leq \alpha$ then at least one of the predictors are useful in predicting the response.

2. Do all of the predictors help to explain Y, or is only a subset of the predictors useful? **Answer**: T-test for each predictor, if p-value is $> \alpha$ then that predictor is not needed in the in model with the presence of the the other predictors.

$H_0: \beta_j = 0$, given $\beta_1, \beta_2, \ldots, \beta_p$ $(i \neq j)$ is in the model.

$H_A: \beta_j \neq 0$

$T = \dfrac{\hat{\beta}_j}{SE_{\hat{\beta}_j}}$ $\quad df = n-p-1$



Answering the Questions

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response? **Answer**: F - test, if *p*-value $\leq \alpha$ then at least one of the predictors are useful in predicting the response.

2. Do all of the predictors help to explain Y, or is only a subset of the predictors useful? **Answer**: T-test for each predictor, if *p*-value is $> \alpha$ then that predictor is not needed in the in model with the presence of the the other predictors.

3. How well does the model fit the data? **Answer**: What is the *RSE* for different models, what is R^2 for different models? Do the plots (residuals, Normal QQ, Standardize Residuals, and Extreme Values) appear to follow the assumptions?

Answering the Questions

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response? **Answer**: F - test, if p-value $\leq \alpha$ then at least one of the predictors are useful in predicting the response.

2. Do all of the predictors help to explain Y, or is only a subset of the predictors useful? **Answer**: T-test for each predictor, if p-value is $> \alpha$ then that predictor is not needed in the in model with the presence of the the other predictors.

3. How well does the model fit the data? **Answer**: What is the *RSE* for different models, what is R^2 for different models? Do the plots (residuals, Normal QQ, Standardize Residuals, and Extreme Values) appear to follow the assumptions?

4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction? **Answer**: Prediction Interval and Confidence Interval.

F-Test: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p$ against H_a : at least one $\beta_j \neq 0$, for $j = 1, 2, \ldots p$. That is at least one predictor could be used in the model.

1. Test statistic: $F = \frac{\overbrace{(SST - SSE)}^{SSR}/p}{SSE/(n-p-1)}$
2. P-value: $P(f_{p,n-p-1} \geq F) \leq \alpha$ we reject the null hypothesis.
3. Output from R last line of summary

```
F-statistic: 59.07 on 3 and 20 DF,  p-value: 4.054e-10
> anova(stock3.lm)
Analysis of Variance Table

Response: Stock_Index_Price
                  Df Sum Sq Mean Sq  F value    Pr(>F)
Interest_Rate      1 894463  894463 172.7117 2.684e-11 ***
Unemployment_Rate  1  22394   22394   4.3241   0.05065 .
Year               1    980     980   0.1892   0.66823
Residuals         20 103579    5179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SST

RHo ∴ At least one β_j is needed in the mode.

SSR = 894463 + 22394 + 980 = 917837

SSE = 103579

$F = \frac{917837/3}{103579/(24-3-1)} = 59.09$

Answering Question 2

T-test: $H_0 : \beta_j = 0$ against $H_a : beta_j \neq 0$ for $j = 1, 2, \ldots, p$, given the other variables are in the model.

1. Test statistic: $t_j = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)}$
2. P-value: $P(t_{n-p-1} \geq |t_j|) \leq \alpha$, we reject the null hypothesis for β_j.
3. Output from R:

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -56523.71  134080.46  -0.422    0.678
Interest_Rate         324.59     123.37   2.631    0.016 *
Unemployment_Rate    -231.48     127.72  -1.812    0.085 .
Year                   28.89      66.42   0.435    0.668
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\therefore$ Since the p-value for testing $\beta_3 = 0$ is very large, year is not significant in predicting stock price.

Model Without *Year*

```
stock2.lm <- lm(Stock_Index_Price~Interest_Rate+Unemployment_Rate,
                data = stock_price)
summary(stock2.lm)

Call:
lm(formula = Stock_Index_Price ~ Interest_Rate + Unemployment_Rate,
data = stock_price)

Residuals:
     Min       1Q   Median       3Q      Max
-158.205  -41.667   -6.248   57.741  118.810

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         1798.4      899.2   2.000  0.05861 .
Interest_Rate        345.5      111.4   3.103  0.00539 **
Unemployment_Rate   -250.1      117.9  -2.121  0.04601 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 70.56 on 21 degrees of freedom
Multiple R-squared:  0.8976,Adjusted R-squared:  0.8879
F-statistic: 92.07 on 2 and 21 DF,  p-value: 4.043e-11
```