Multiple Linear Regression

Links: MATH 4322

(Machine Learning Lecture 5; Textbook Sections 3.2 & 6.1)

Stock Price 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`

(data can be found in the `stock_price.csv` in canvas)

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?

We say a model is good at predicting the response (output) by quantifying how well the model fits the data. The two quantities we use for this are residual standard error (RSE) and the coefficient of determination (R^2). In *R*, these quantities are in the `summary` output of the `lm()` function.

Stock Price Example with Interest Rate as Predictor

The Estimate of the Simple Linear Regression Model

$y \sim x$, data =

```
> stock.lm <- lm(Stock_Index_Price~Interest_Rate,data = stock_price)
> summary(stock.lm)

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

Residuals:
Min        1Q    Median       3Q      Max
-183.892  -30.181    4.455   56.608  101.057

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)      -99.46    95.21  -1.045     0.308
Interest_Rate    564.20    45.32  12.450 1.95e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 75.96 on 22 degrees of freedom
Multiple R-squared:  0.8757,Adjusted R-squared:  0.8701
F-statistic:   155 on 1 and 22 DF,  p-value: 1.954e-11
```

Residuals: $y_i - \hat{y}_i$

$-99.46 = \hat{\beta}_0$

$564.20 = \hat{\beta}_1$

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

$H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$

Equation: $\widehat{\text{stock\_index\_price}} = -99.46 + 564.2 * \text{Interest\_Rate}$

In this case $R^2 = 0.8757$, so about 87.5% of the variation in the stock index price can be explained by the equation.

(further recall Linear Regression > Assumptions about the Model)

Why We Don't Do Separate 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 *j*th predictor
- β_j quantifies the association between the *j*th 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: $st\_id\_pr = \beta_0 + \beta_1 \times \ Interest\_Rate + \beta_2 \times \ Unemployment\_Rate + \beta_3 \times \ Year \ +$ (`st_id_pr` means `stock_index_price` here, I shortened it so it could fit on the screen)

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 *i*th 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

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

(Recall: Linear Regression > Principle of Least Squares)

Linear Model of the Stock Index Price Example

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:                  β0
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -56523.71  134080.46  -0.422    0.678
Interest_Rate         324.59 = β1 123.37   2.631    0.016 *
Unemployment_Rate    -231.48 = β2 127.72  -1.812    0.085 .
Year                  28.89 = β3  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
```

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

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 $\hat{\beta}_3$.

```
#Interpret B_3: For each additional year the stock index price increases
#on average by $28.89 , given a fixed interest rate and unemployment rate.
```

(the reason it says 1% is because that is the unit of `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
```

(When we add more variables, we look at *Adjusted R-squared* instead of the multiple R-squared).

Important Questions for Multivariate Regression

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

- Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
  - answer: Perform the F-test, if the *p-value* $< \alpha$ then at least one of the predictors are useful in predicting the response.
  - (in this case the degrees of freedom is $n - p - 1$)
- Do all of the predictors help *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 model with the presence of the other predictors.
- How well does the model fit the data?
  - answer: What is the RSE for the different models, what is the Coefficient of Variation (R^2) for the different models? Do the plots (residuals, Normal QQ, Standardize Residuals, and Extreme Values) appear to follow the assumptions?
  - (there are 5 statistics we will look at to determine the best model)
- Given a set of predictor values, what response value should we predict and how accurate is our prediction?

- <u>answer:</u> Prediction Interval and Confidence Interval.

Stock Price Example: Answering Questions 1 & 2

(Recall the [full model for the stock price example])

Answering Question 1

F-Test: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ 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: $F = \frac{(SST - SSE)/p}{SSE/(n-p-1)}$
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
```

SSR

SSE

1021416 = SST

$$F = \frac{(1021416 - 103579)/3}{103579/(24-3-1)} = 59.07$$

P-value $= P(F \geq 59.07) = 1 - \text{pf}(59.07, 3, 20) \approx 0$ RHo

Cathy Poliak, Ph.D. cpoliak@central.uh.edu Sections 3.2 & 6.1 21/26

At least one β_j's are not zero.

(the 1021416 is the SST, and the 103579 is the SSE; see [Linear Regression > Calculating $R 2$])

In this case you can see from the last line of summary output that the p-value of the F-statistic is less than 0.05, this means that at least one of the predictors are useful in predicting the response (more formally: At least one β_j's are not zero).

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: $t_j = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)}$
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: df

```
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
```

$\leq 2*P(T \leq -2.63)$
$= 2*pt(-2.631, 20)$

$H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$ $t = 2.631$, p-value$=0.016$ R H_0.

Interest rate can be used as a predictor if unemployment rate and year are in the model.

(On the very right of the summary output you can see the p-values for the t-test (these are two tailed)).
degrees of freedom: $n - p - 1$
(cool thing in R: if you see a `*` or a `.` next to the p value numbers in the list, that means you probably want to keep those and not include the others).

Based on this, we can probably drop `year` from the model and only keep the other two predictors.

Choosing the Best Predictors

- We look at the individual p-values the lower the p-value the more significant the predictor is used in the model. We can remove the predictors that have higher p-values.
- **Problem:** This p-value is calculated given that all of the other predictors are in the model thus if the number of predictors are large we are likely to make some false discoveries.
- Thus we have to look at all possible models to determine which model works best. **Problem:** there are 2^p models that contain subsets of p variables (predictors) (see Subsets & Power Sets > Power Set)

- The **stepwise** regression (or stepwise selection) consists of iteratively adding and removing predictors, in the predictive model, in order to find the subset of variables in the data set resulting in the best performing model, that is a model that lowers prediction error.

Stepwise Regression

There are three strategies of stepwise regression

1. **Forward** selection, which starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant.
2. **Backward** selection (or backward elimination), which starts with all predictors in the model (full model), iteratively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant.
3. **Mixed** selection (or sequential replacement), which is a combination of forward and backward selections. You start with no predictors, then sequentially add the most contributive predictors (like forward selection). After adding each new variable, remove any variables that no longer provide an improvement in the model fit (like backward selection).

In R this is done with `step(name.lm)` (for backwards)

- `step(name.lm, direction = "forward")`
- `step(name.lm, direction = "both")`

example:

```
> step(stock3.lm) #stepwise "backwards"
Start:  AIC=208.88
Stock_Index_Price ~ Interest_Rate + Unemployment_Rate + Year

                    Df Sum of Sq    RSS    AIC
- Year               1       980 104559 207.11
<none>                           103579 208.88
- Unemployment_Rate  1     17012 120591 210.53
- Interest_Rate      1     35847 139426 214.01

Step:  AIC=207.11
Stock_Index_Price ~ Interest_Rate + Unemployment_Rate

                    Df Sum of Sq    RSS    AIC
<none>                           104559 207.11
- Unemployment_Rate  1     22394 126953 209.76
- Interest_Rate      1     47932 152491 214.16

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

Coefficients:
(Intercept)      Interest_Rate  Unemployment_Rate
1798.4              345.5             -250.1
```

Lecture TL;DR

We need to put multiple predictors in the model, and need to check if the model is good. We also need to check which predictors we actually need, so we use things like stepwise regressions to choose a good subset of the model to use.