Multiple Linear Regression

Section 3.2 & 6.1

Dr. Cathy Poliak, cpoliak@uh.edu

University of Houston

Continuing 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 Canvas This is from https://datatofish.com/multiple-linear-regression-in-r/

Questions We Want To Answer

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.

Calucations Used to Answer These Questions

1. Residual sum of squares:

$$SSE = RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ variance of residuals

2. Sum of squares regression:

$$= SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$ variance of the predicted value

3. Total sum of squares:

$$SST = TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$ variance of the observed response values

TSS = SSR + SSE

Then the $F-$statistic is calculated by:

$$F = \frac{SSR/p}{RSS/(n-p-1)} = \frac{MSR}{MSE}$$

Putting these Values in a Table: **ANOVA**

"**Analysis of Variance** (ANOVA) table consist of calculations that provide information about levels of variability within a regression model and form a basis for tests of significance." [1]

Source	Df	Sum Sq	Mean Sq	F-value	P-value
Model	p	SSR	$\frac{SSR}{p} = MSR$	$\frac{MSR}{MSE}$	$P(f_{p,n-p-1} \geq F)$
Residuals	$n-p-1$	RSS	$\frac{RSS}{n-p-1} = MSE$		
Total	$n-1$	TSS			

Note: SSR is the total variation accounted in the model among all of the p predictors. `R` separates this by each p predictor

[1] http://www.stat.yale.edu/Courses/1997-98/101/anovareg.htm

```
stock3.lm <- lm(Stock_Index_Price~Interest_Rate+Unemployment_Rate+Year,
                data = stock_price)
(stock3.aov = 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
Total             23 1021416
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$n=24$

$H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$ if β_2 & β_3 are in the model

1. $SSR = 894463 + 22394.17 + 979.9 = 917837.1$
2. $RSS = 103578.7$
3. $TSS = SSR + RSS = 1021416$

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

H_A: At least one $\beta_j \neq 0$

$$F = \frac{SSR/p}{RSS/(n-p-1)} = \frac{917837.1/3}{103578.7/(24-3-1)} = 59.07$$

p-value = 1 - pf(59.07, 3, 20) ≈ 0

Without Year

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

```
Analysis of Variance Table

Response: Stock_Index_Price
                  Df Sum Sq Mean Sq  F value    Pr(>F)    
Interest_Rate      1 894463  894463 179.6477 9.231e-12 ***
Unemployment_Rate  1  22394   22394   4.4977   0.04601 *  
Residuals         21 104559    4979                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR = 894463 + 22394 = 916857$

$RSS = 104559$

$TSS = 104559 + 916857 = 1021416$

$$F = \frac{916857/2}{104559/(24-2-1)} = 92.07$$

Answering Question 3: Common Numerical Measures of the Model Fit

$= SSR/TSS = \frac{TSS - RSS}{TSS}$

1. $R^2 = 1 - \frac{RSS}{TSS}$ This the the fraction of the variability in Y that can be explained by the equation. We desire this to be close to 1.
2. Residual Standard Error = RSE = $\sqrt{\frac{RSS}{n-p-1}}$, the variability of the residuals. We desire this to be small.
3. **Problem**: as we add more variables, the R^2 will increase.
4. We have a number of techniques for adjusting to the fact that we have more variables.

Compare Values

Predictors	RSE	R^2
Interest_Rate + Unemployment_Rate + Year	71.96	0.8986
Interest_Rate + Unemployment_Rate	70.56	0.8976
Interest_Rate	75.96	0.8757

Other Statistics to Choose Best Linear Model

We can then select the best model out of all of the models that we have considered. How do we determine which model is best? Various statistics can be used to judge the quality of a model.

These include:

- *Mallows'* C_p,
- *Akaike information criterion* (*AIC*),
- *Bayesian information criterion* (*BIC*) and
- *adjusted* R^2.

We desire a model with small values of C_p, *AIC*, and *BIC* and large (close to 1) *adjusted* R^2.

Adjusted R^2

- As stated before, the problem is that the more predictors we drop the from the model the R^2 becomes lower.
- For a least squares model with p variables, the adjusted R^2 is calculated as
$$1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}$$
- We desire again a large adjusted R^2.

From the summary output

Multiple R-squared: 0.8986, $\boxed{\text{Adjusted R-squared: 0.8834}}$

Adjusted R^2 Calculations

SST $= 1021416$

Predictors	RSS	Adj. R^2
Interest_Rate + Unemployment_Rate + Year	103579	$1 - \frac{103579/(24-3-1)}{1021416/23} = 0.8834$
Interest_Rate + Unemployment_Rate	104559	? 0.8879
Interest_Rate	126953	$1 - \frac{126953/(24-1-1)}{1021416/23} = 0.8701$

1. Determine the adjusted R^2 for the model with the 2 predictors.

   a) 104559
   b) 1021416
   c) 0.8976
   d) 0.8879

$$\text{adj. } R^2 = 1 - \frac{104559/(24-2-1)}{1021416/23} = 0.8879$$

C_p

- Mallows' C_p compares the precision and bias of the full model to models with a subset of the predictors.
- Usually, you should look for models where Mallows' C_p is small and close to the number of predictors in the model plus the constant $(p+1)$.
- A small Mallows' C_p value indicates that the model is relatively precise (has small variance) in estimating the true regression coefficients and predicting future responses.
- A Mallows' C_p value that is close to the number of predictors plus the constant indicates that the model is relatively unbiased in estimating the true regression coefficients and predicting future responses.
- Models with lack-of-fit and bias have values of Mallows' C_p larger than p.
- Formula for C_p:

$$C_p = \frac{\text{RSS}_p}{\text{MSE}_{\text{all}}} + 2(p+1) - n$$

Where p is the number of predictors in the model and RSS_p is the residual sum of squares from the model with p predictors and MSE_{all} is the MSE for the model with all the predictors.

Stock Price Example

Output from model:
$Stock\_Index\_Price = \beta_0 + \beta_1 \times Interest\_Rate + \beta_2 \times Unemployment\_Rate + \beta_3 \times Year + \epsilon$

```
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
```

$\boxed{5179} = 103579/20$

$MSE_{All} = 5179 \quad RSS_3 = 103579$

$$C_3 = \frac{103579}{5179} + 2(3+1) - 24 = 3.9998 \approx 4 = p+1$$

Output from model: $Stock\_Index\_Price = \beta_0 + \beta_1 \times Interest\_Rate + \epsilon$

```
stock.lm = lm(Stock_Index_Price ~ Interest_Rate,data = stock_price)
anova(stock.lm)
```

```
Analysis of Variance Table

Response: Stock_Index_Price
              Df Sum Sq Mean Sq F value    Pr(>F)
Interest_Rate  1 894463  894463     155 1.954e-11 ***
Residuals     22 126953    5771
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$RSS_1 = 126953$

$MSE_{All} = 5179$

$$C_p = \frac{126953}{5179} + 2(1+1) - 24 = 4.513$$

Lab Question

The following is an output for the model:

$$Stock\_Index\_Price = \beta_0 + \beta_1 \times Interest\_Rate + \beta_2 \times Unemployment\_Rate + \epsilon$$

```
anova(stock2.lm)
```

```
Analysis of Variance Table

Response: Stock_Index_Price
                  Df Sum Sq Mean Sq  F value    Pr(>F)
Interest_Rate      1 894463  894463 179.6477 9.231e-12 ***
Unemployment_Rate  1  22394   22394   4.4977   0.04601 *
Residuals         21 104559    4979
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. Determine the C_p statistic.

$$C_2 = \frac{104559}{5179} + 2(2+1) - 24$$

a) 2

b) 104559

c) 2.189

d) 4.513

AIC

- **Akaike information criterion** (AIC) is an estimator of the relative quality of statistical models for a given set of data.
- Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models.
- AIC is used in the step() function in R and provides a means for model selection. The default is the "backward" selection process.
- The calculation is for p variables:

$$2(p+1) + n \ln\left(\frac{\text{RSS}}{n}\right)$$

- The smaller the AIC the better the fit.

AIC Calculations

Predictors	RSS	AIC
Interest_Rate + Unemployment_Rate + Year	103579	$2(4)+24*\ln\left(\frac{103579}{24}\right)=208.88$
Interest_Rate + Unemployment_Rate	104559	? 207.11
Interest_Rate	126953	$2(2)+24*\ln\left(\frac{126953}{24}\right)=209.76$

3. Determine the AIC for the model with the 2 predictors.

a) 207.11
b) 203.11
c) 104559
d) 4356.625

$$AIC = 2(3) + 24 * \ln\left(\frac{104559}{24}\right)$$

From the step() Function

```
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
```

BIC

- Derived from a Bayesian point of view. Call the Schwartz's information criterion.
- Similar to the AIC and C_p.
- We generally select the model with the lowest BIC value.
- Formula
$$BIC = -2 * loglikelihood + log(n)(p+1)$$
- There are several ways to estimate this value. In R we can use the function `BIC`

```
BIC(stock.lm) #Interest_Rate
```

```
[1] 283.4076
```

```
BIC(stock2.lm) #Interest_Rate + Unemployment_Rate
```

```
[1] 281.9281
```

```
BIC(stock3.lm) #Interest_Rate + Unemployment_Rate + Year
```

```
[1] 284.8801
```

Which Subsets of Parameters are Best?

Predictors	R^2	Adj. R^2	C_p	AIC	BIC
Interest_Rate + Unemployment_Rate + Year	0.8986	0.8834	4.0	208.88	284.8801
Interest_Rate + Unemployment_Rate	0.8976	0.8879	2.1892	207.11	281.9281
Interest_Rate	0.8757	0.8701	4.5133	209.76	283.4076

4. According to these statistics which model is best?
   a. With Interest Rate only
   b. With Interest Rate and Unemployment Rate
   c. With all three predictors
   d. Any of these models will be fine

Function to Get Best Subset

- The regsubsets() function (part of the leaps library) performs best subset selection by identifying the best models that contains a given number of predictors.
- The *best* is quantified using the RSS.
- The syntax is the same as for lm().
- Type in the following and run in R.

```
library(leaps)
stock.fit = regsubsets(Stock_Index_Price~Unemployment_Rate +
                                   Interest_Rate + Year,
                                   data = stock_price)
(stock.res = summary(stock.fit))
```

- An asterisk indicates that a given variable is included in the corresponding model. For instance, this output indicates that the best one-variable model contains Interest_Rate.
- The summary() function also returns R^2, SSR, adjusted R^2, C_p, and and estimated BIC.

```
Subset selection object
Call: regsubsets.formula(Stock_Index_Price ~ Unemployment_Rate + Interest_R
    Year, data = stock_price)
3 Variables  (and intercept)
                  Forced in Forced out
Unemployment_Rate     FALSE      FALSE
Interest_Rate         FALSE      FALSE
Year                  FALSE      FALSE
1 subsets of each size up to 3
Selection Algorithm: exhaustive
         Unemployment_Rate Interest_Rate Year
1  ( 1 ) " "               "*"           " "
2  ( 1 ) "*"               "*"           " "
3  ( 1 ) "*"               "*"           "*"
```

Show The Statistics From the `regsubests()`

```
stock.stat = cbind(stock.res$rsq,
                   stock.res$adjr2,
                   stock.res$cp,
                   stock.res$bic)
colnames(stock.stat) = c("rsq","Adjr2","Cp","BIC")
stock.stat
```

5. Which of the following statistic do we want the highest value?

   a) adjusted R^2
   b) C_p
   c) BIC
   d) AIC

	rsq	Adjr2	Cp	BIC
Interest_Rate	0.8757090	0.8700594	4.513301	-43.68700
Interest_Rate + Unemployement_Rate	0.8976336	0.8878844	2.189215	-45.16656
Interest_Rate + Unemployment_Rate + Year	0.8985930	0.8833819	4.000000	-42.21449

Assumptions about the Model

The linear regression model has assumptions that we need to prove is true. We use the acronym **LINE** to remember these assumptions.

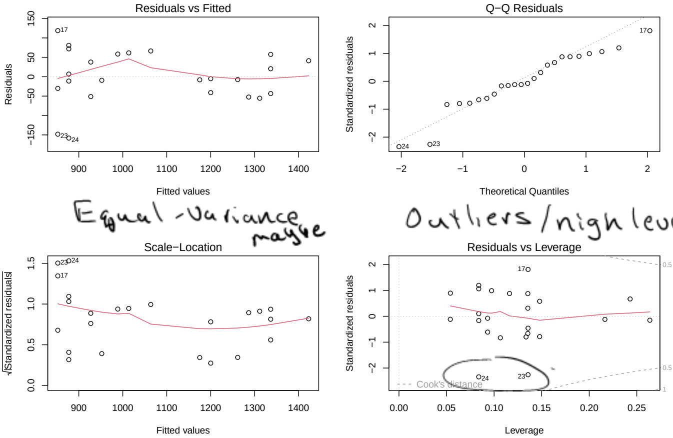
- **L**inear relationship: can we determine a linear relationship between the response an other variables?
- **I**ndependent observations: are the observations a result of a simple random sample?
- **N**ormal distribution: for any fixed value of X, Y is normally distributed.
- **E**qual variance: the variance of the residual is the same for any value of X.
- Be careful of extreme values.

Diagnostic Plots to Check Assumptions

Equation:
$\widehat{\text{Stock\_index\_price}} = 1798.4 + 345.5 \times \text{Interest\_Rate} - 250.1 \times \text{Unemployement\_Rate}$

Linear maybe   Normal ✓

Equal-Variance maybe   Outliers/high leverage



Answering Question 4: Predictions

Recall we want to make a prediction on $Y = f(X) + \epsilon$. We found an estimate for $f(X)$:

$$f(\hat{X}) = \text{Stock\_in}\hat{\text{d}}\text{ex\_price} = 1798.4 + 345.5 \times \text{Interest\_Rate} - 250.1 \times \text{Unemployement\_Rate}$$

What is the predicted value of the `stock index price` if Interest_Rate $= 2.25$ and Unemployment_Rate $= 6.0$?

$$\widehat{\text{stock\_index\_price}} = 179.4 + 345.5 * 2.25 - 250.1 * 6.0$$
$$= 1075$$

Reducible and Irreducible Error $Y = f(X) + \varepsilon$

There are uncertainty associated with this prediction.

1. The coefficients are only an estimate for the true population model $f(X)$. This is related to the **reducible error**. We use the **confidence interval** for the predicted value to determine how close $\hat{Y}$ will be to $f(X)$.

2. We are assuming a linear model for $f(X)$, so there is an additional source of potentially reducible error which we call *model bias*.

3. Even if we know $f(X)$, the response value cannot be predicted perfectly because of the random error ε. This is the **irreducible error**. How much will Y vary from $\hat{Y}$? We use **prediction intervals** to answer this question.

Confidence Interval

```
predict(stock2.lm,
        newdata = data.frame(Interest_Rate = 2.25,
                             Unemployment_Rate = 6.0),
        interval = "c")
```

$\hat{y}$

```
      fit      lwr      upr
1 1074.99 975.9122 1174.067
```

This means we predict the **average** stock index price among all of the months with 2.25% interest rate and 6% unemployment to be between [975.9122, 1174.067] with 95% confidence.

Prediction Interval

```
predict(stock2.lm,
        newdata = data.frame(Interest_Rate = 2.25,
                             Unemployment_Rate = 6.0),
        interval = "p")
```

$\hat{y}$

```
      fit     lwr      upr
1 1074.99 897.932 1252.047
```

This means the predicted stock index price for a particular month with 2.25% interest rate and 6% unemployment rate is between [897.932,1252.047] with 95% confidence.

This interval is wider than the confidence interval, because it incorporates both the error in the estimate for $f(X)$ (*reducible error*) and the uncertainty as to how much an individual point will differ from the population model (*irreducible error*).