Other Considerations in the Regression Model

Section 3.3 & 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?

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
```

$H_0: \beta_j = 0 \quad H_A: \beta_j \neq 0$, given $\beta_1, \ldots, \beta_p$ is in the model

$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \quad H_A$: At least one $\beta_j \neq 0$

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

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.

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
```

Model: $\hat{y} = 1798.4 + 345.5 * \text{Interest\_rate} - 250.1 * \text{Unemployment\_Rate}$

```
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
```

Answering Question 3: Common Numerical Measures of the Model Fit

1. R^2 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. RSE = Residual Standard Error, 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

$$R^2 = 1 - \frac{SSE}{SST}$$

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

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

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

$$RSE = \sqrt{\frac{SSE}{n-p-1}}$$

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*.

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.

Calculation of C_p

Given the ANOVA Table:

	Df	Sum Sq	Mean Sq	F	P-value
Regression	p	SSR	MSR = $\frac{\text{SSR}}{p}$	$\frac{\text{MSR}}{\text{MSE}}$	$p - value$
Residuals	$n-p-1$	SSE	MSE = $\frac{\text{SSE}}{n-p-1}$		
Total	$n-1$	SST			

Formula for C_p:

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

Where p is the number of predictors in the model and SSE_p is the SSE 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* $+ \epsilon$

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

$= \frac{SSE}{df = n-p-1}$

$C_3 = \frac{SSE_3}{MSE_3} + 2(3+1) - 24 = \frac{103579}{5179} + 8 - 24 = 4$

$n-p-1 + 2(p+1) - n = p+1$

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

```
> stock.lm <- lm(Stock_Index_Price~Interest_Rate)
> 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
```

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

Lab Question

1. The following is an output for the model:
   $Stock\_Index\_Price = \beta_0 + \beta_1 \times Interest\_Rate + \beta_2 \times Unemployment\_Rate + \epsilon$

```
> 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
```

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

Determine the C_p statistic.

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:

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

- The smaller the AIC the better the fit.

AIC Calculations

Predictors	SSE	AIC
Interest_Rate + Unemployment_Rate + Year	103579	$2(4) + 24 * \ln\left(\frac{103579}{24}\right) = 208.88$
Interest_Rate + Unemployment_Rate	104559	$2(3) + 24 * \ln\left(\frac{104559}{24}\right) = 207.11$
Interest_Rate	126953	$2(2) + 24 * \ln\left(\frac{126953}{24}\right) = 209.76$

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

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

From the step() Function

```
> step(stock3.lm)
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 + Unemployement_Rate + Year
[1] 284.8801
```

Adjusted R^2

- Recall the usual $R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$
- 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
$$1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}$$
- We desire again a large adjusted R^2.

Adjusted R^2 Calculations

SST = 1021416

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

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

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

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

Total number of possible models $2^3 = 8$

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 SSE.
- 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)
stock.res
```

- 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_Rate + 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 ) "*"               "*"           "*"
```

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * interest\_rate$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * int\_rate + \hat{\beta}_2 * un\_rate$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * in\_rate + \hat{\beta}_2 * unep\_rate + \hat{\beta}_3 * year$

Show The Statistics From the `regsubests()` Function

```
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
```

4. 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
[1,] 0.8757090 0.8700594 4.513301 -43.68700
[2,] 0.8976336 0.8878844 2.189215 -45.16656
[3,] 0.8985930 0.8833819 4.000000 -42.21449
```

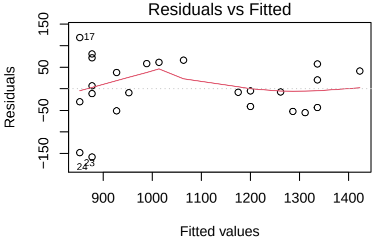
$\therefore$ The best model is:

$$\widehat{stock\_index\_price} = \hat{\beta}_0 + \hat{\beta}_1 * interest\_rate + \hat{\beta}_2 * unemployment\_rate$$

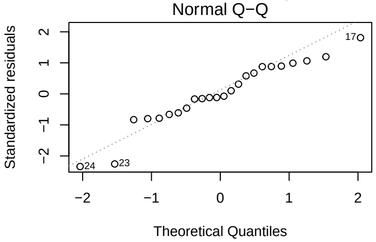
Answering Question 3

LINE

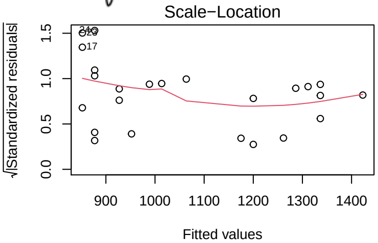
Linear - No form



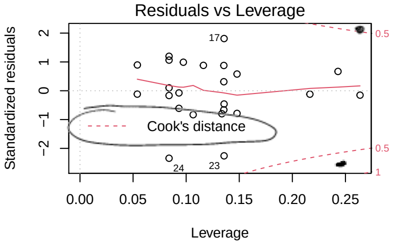
Normal - follow 45° Line



Equal Variance - No form



Outliers



Answering Question 4: Intervals for Regression

- Prediction interval

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

Answering Question 4: Intervals for Regression

- Prediction interval

```
> predict(stock2.lm,
+         newdata = data.frame(Interest_Rate = 2.25,
+                                              Unemployment_Rate = 6.0),
+         interval = "p")
      fit     lwr      upr
1 1074.99 897.932 1252.047
```

$897.932 \leq y_i \leq 1252.047$

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.

- Confidence interval

```
> #Confidence Interval
> predict(stock2.lm,
+         newdata = data.frame(Interest_Rate = 2.25,
+                                   Unemployment_Rate = 6.0),
+         interval = "c")
      fit      lwr      upr
1 1074.99 975.9122 1174.067
```

$975.9122 \leq \mu_y \leq 1174.067$

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.

Lab Question

The following are intervals for the stock index price where unemployment rate is at 5.5% and interest rate is at 3% with 95% confidence:

Prediction Interval: [1244.12, 1674.31]

Confidence Interval: [1301.96, 1616.48]

5. Which interval predicts the stock index price for **one** observation?

    a) Prediction interval
    b) Confidence interval