Regression Model - Best Subset & Regression Intervals

Links: MATH 4322

Recall from the previous note

ⓘ **Stock Price Example & its Linear Model** ›

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:

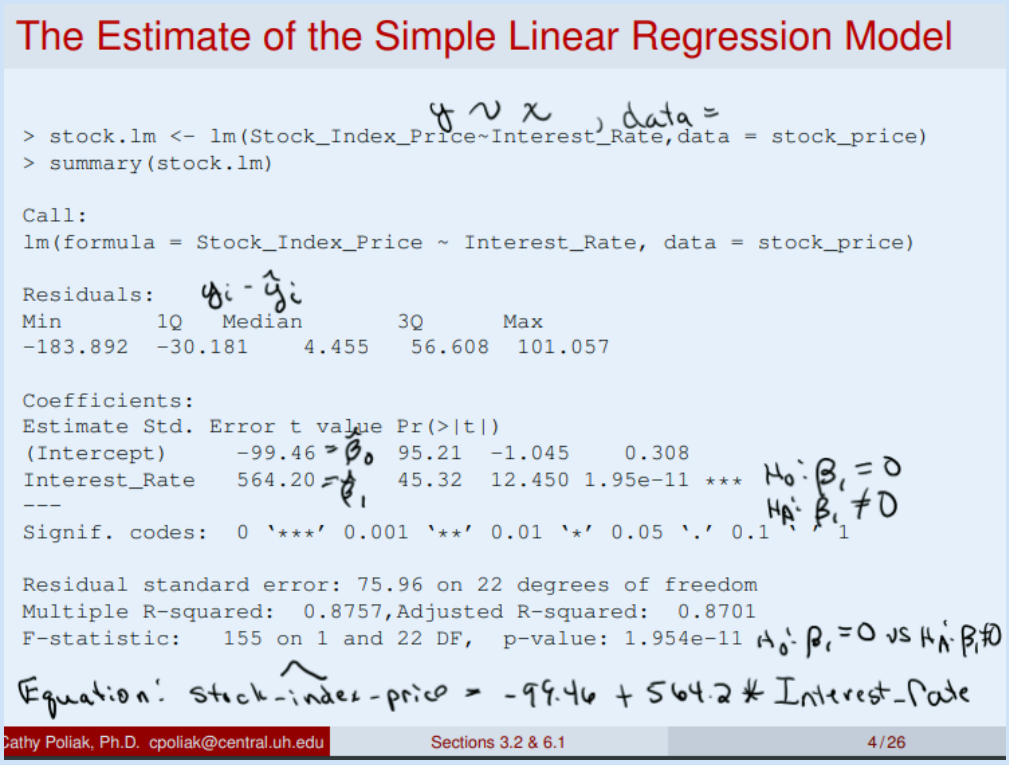
- `Interest_Rate`
- `Unemployment_Rate`
- `Year`

(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 ›



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)

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:
                  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
```

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

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).

ⓘ Questions to answer for Multivariate Regression ›

Important Questions for Multivariate Regression

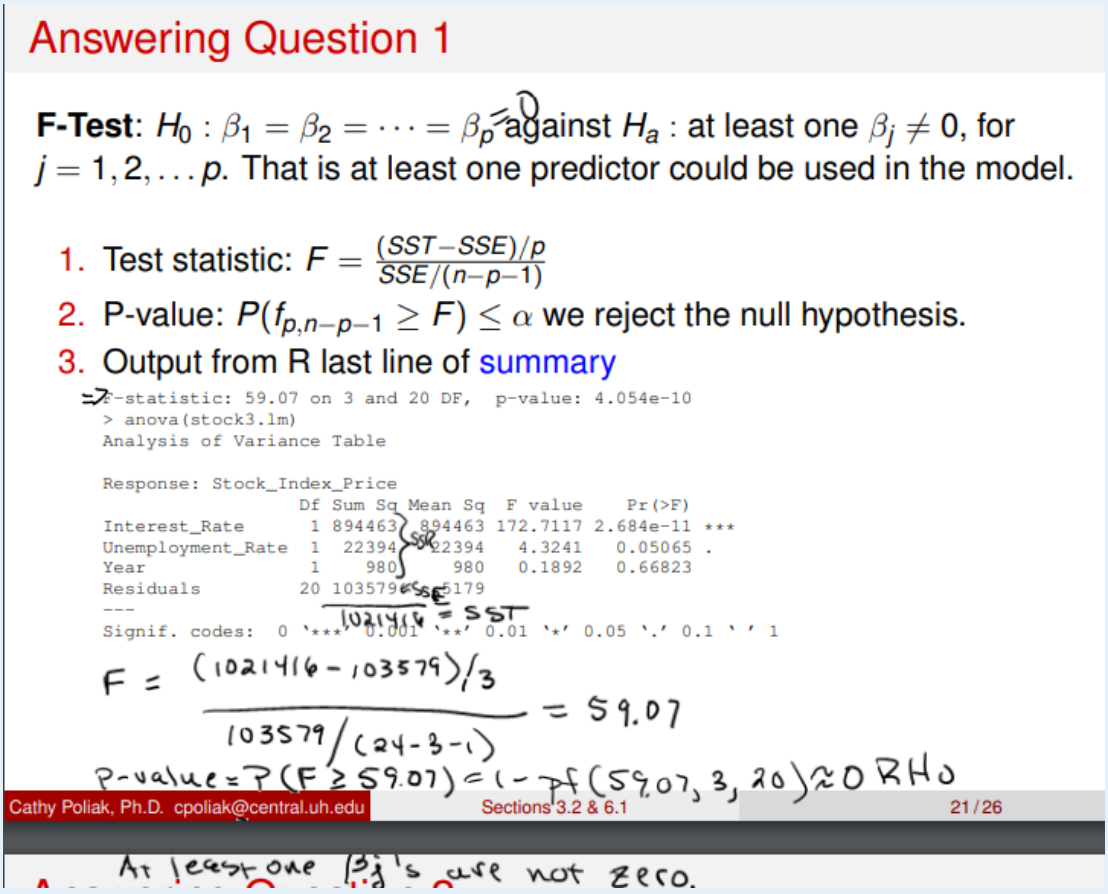
For the **multivariate regression** we are interested in answering a few
important questions.

- 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?
    - answer: Prediction Interval and Confidence Interval.

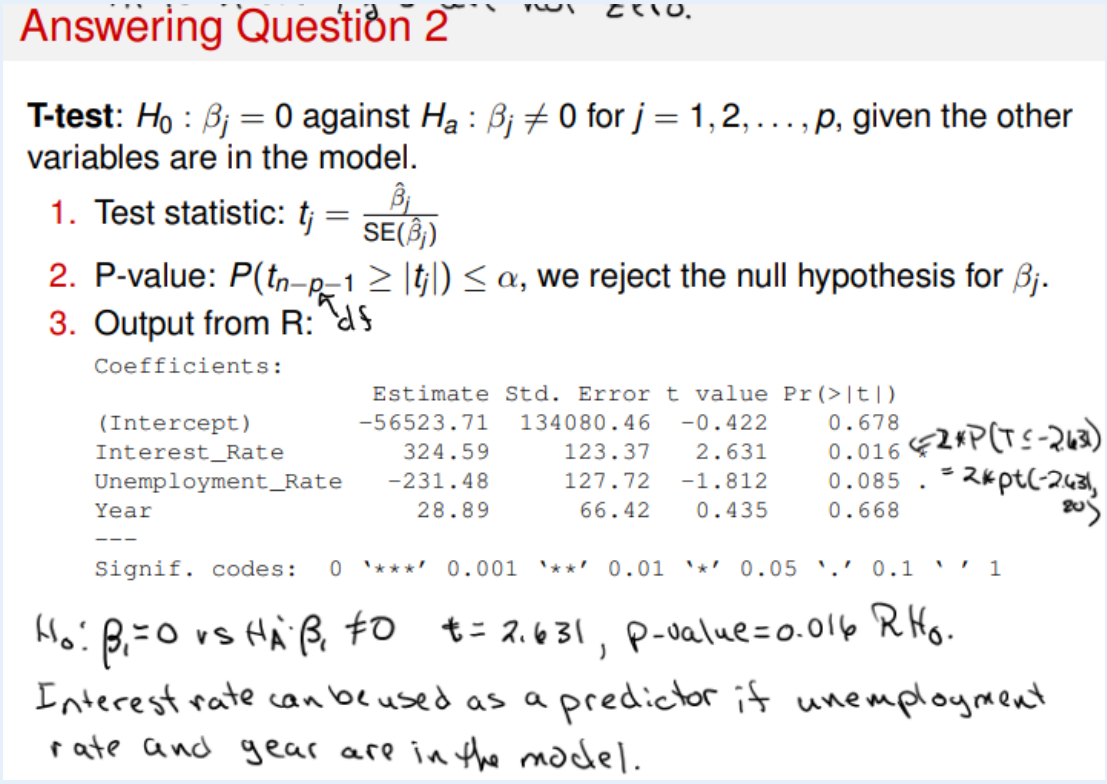
Stock Price Example: Answering Questions 1 & 2

(Recall the full model for the stock price example)



(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: $t_j = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)}$
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
```

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

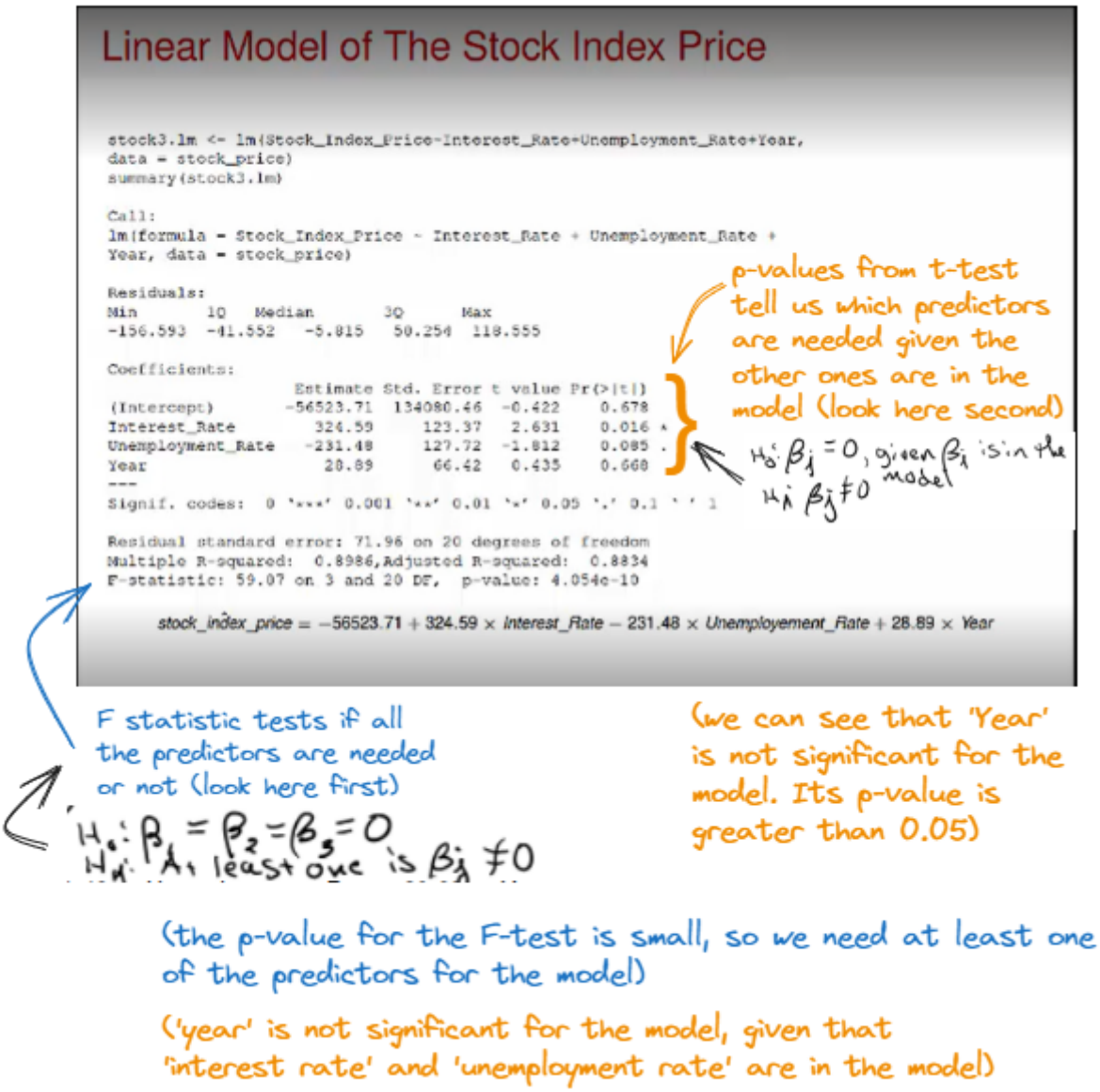
Interest rate can be used as a predictor if unemployment rate and year are in the model.

(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.

Answers to Question 1 & 2

(Questions 1 & 2: is at least one predictor useful in predicting the response? Which predictors actually help?)



Other values to look at are the Residual standard error (which we want to be low), and the **adjusted** R-squared (which we want to be as close to one as possible). (Why we look at adjusted r-squared instead of the other one will be explained in this note).

A good model is one with a small RSE, and a large (close to one) R^2. This model is good, but there are other models with (one or more of) these predictors that could potentially be used, as explained previously, there are 2^p possible models to choose from (so in this case, with three predictors there are 8 possible models that could be used). We have ways to automate good model selection (see Multiple Linear Regression > Stepwise Regression).

Best Subset of Predictors

We determined previously that `year` could be removed as a predictor;

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
```

Notice that the *p-value* for the intercept went down dramatically ("this means that removing the year was probably a good move" - Prof. Poliak).

The *p-value* for the F-statistic is small so that means at least one of the predictors is significant, and all the *p-values* for the the t-test for the predictors show that they are significant.

Answering Question 3: Common Numerical Measures of the Model Fit

(less fancy terms: we're trying to figure out how well the model fits the data)

Recall that R^2 is the fraction of variability in *Y* that can be explained by the equation, we want this to be close to 1. The RSE is the variability in residuals, we want this to be small. (See R-squared, Residual Standard Error, and RSE & R-squared).

The Problem here is that (Multiple) R-squared increases as we add more predictors (that is why we look at Adjusted R-squared). We have a number of techniques for adjusting to the fact that we have more variables.

(the values could be standardized, but we don't do that since we want to interpret the coefficients of those values).

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 (the *p* is the number of predictors)
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)
- Adjusted R^2

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

Mallows' C_p

- Mallows' C_p compares the precision and bias of the full model to models with a subset of the predictors (what its doing is looking at the Sum Squares Error for each reduced model and dividing it by the Mean Squared Error of the full model).
- 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*.

The calculation of the C_p can come from the ANOVA Table in `R`, from the slides:

Calculation of C_p

Given the ANOVA Table:

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

Formula for C_p:

$$C_p = \frac{SSE_p}{MSE_{all}} + 2(p+1) - n$$

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.

Example with Stock Price Dataset

for the full model:

```
# (import stock price data set found in canvas)
# full model, with all 3 predictors
stock3.lm <- lm(Stock_Index_Price~Interest_Rate+

Unemployment_Rate+

Year, data = stock_price)
anova(stock3.lm)
```

Running `anova` on the full model you will see that the Sum Squares (SSE) for the full model is *103579*, and the Mean Squared Error (MSE) for the full

model is *5179* (look at the last row where it says "Residuals", that is the total where you find these values).

So the Mallows' C_p for the full model is

$$C_3 = \frac{103579}{5179} + 2(3+1) - 24 = 3.999807 \approx 4$$

(you will see in the dataset that the number of observations is 24, so $n = 24$).
(in R you can do this with: `(103579/5179)+2*(3+1)-24`, its not exactly 4 because of rounding errors). You can see for the full model its the same as the number of predictors plus one.

for only interest rate as the predictor:

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

You will see that the SSE for this model is 126953 (we calculated the MSE for the full model earlier, *we only use the MSE for the full model with all the predictors!*)

The Mallows' C_p for this model is

$$C_1 = \frac{126593}{5179} + 2(1+1) - 24 = 4.513$$

"So its higher than the 4 (the C_p of the full model), so maybe this isn't that good." - Prof. Poliak

for Interest Rate & Unemployment Rate as the predictors

```
stock2.lm <- lm(Stock_Index_Price~Interest_Rate+

Unemployment_Rate,
                                                    data =
```

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

From the `anova` output we will see that the SSE for this model is 104559 (again remember, for the MSE we only use it from the full model!)

The C_p statistic for this model is

$$C_2 = \frac{104559}{5179} + 2(2+1) - 24 = 2.189033$$

"This is probably the better model, and this C_p confirms this" - Prof. Poliak
(Yes, you may have to calculate this on the test)

Akaike Information Criterion (AIC)

The 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 in *R* and provides a means for model selection. The default is "backward" selection process.

The calculation is for *p* variables:

$$2(p+1) + n\ ln(\frac{SSE_p}{n})$$

The smaller the AIC the better the fit.
(you can get the *SSE* of that model from its `anova` table).

In the step function in R, the predictor names on the left indicate which ones were removed.

(if you do the AIC for the stock index model with only interest rate and unemployment rate as predictors you will see that will be the smallest AIC of all the models we have tried so far. So that model is so far looking to be the best one).

Bayesian Information Criterion (BIC)

Derived from a Bayesian point of view, called the Schwartz's information criterion. It is similar to the AIC and the Mallows Cp.

We generally select the model with the lowest BIC value, the formula for it is

$$BIC = -2 * loglikelihood + log(n)(p+1)$$

There are several ways to estimate this value, in R we can use the function `BIC`.

"I am going to hand wave over this one... there are many ways to get loglikelihood... you don't have to do the calculation" -- Poliak

Adjusted R^2

(Recall the usual way to calculate R-squared)

The problem is that more predictors we drop from the model the R^2 becomes lower. For a least squares with *p* variables, the adjusted R^2 is calculated as

$$1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

We desire again a large adjusted R^2. (Recall that SST (sum squares total) can be calculated as the sum of regression sum squares and sum squares error, you can get both of those in the `anova` output).

(Stock Price Example) 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

(remember these are only 3 models, there a total of 8 possible ones)

Function to Get Best Subset

The `regsubsets()` function (part of the `leaps` library) performs best subset by identifying the best models that contains a given number of predictors. The *best* is quantified using SSE.
The syntax is the same as for `lm()`.

example 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, in this output that the best one-variable model contains `Interest_Rate`. (Basically, if it has a star that means those are the best variables to use for the amount of predictors specified). This function only goes up to 8 as a default.

The `summary()` output also returns R^2, SSR, Adjusted R-squared, Mallows' Cp, and an estimated BIC.

You can display some of those four statistics from the `regsubsets` as follows (this is optional)

```
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
```

```
# the stock.res$rsq doesn't help since that's regular R-squared
# look at the adjusted one instead (see above sections)
```

(Actually) Answering Question 3

Recall the LINE assumptions of the linear model.

For the stock price example, all of our previous statistics shows that the two predictor model is likely the best one.
Remember that is:

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

Plot the diagnostic plots to check the assumptions with the following:

```
par(mfrow = c(2,2))
plot(stock2.lm)
par(mfrow = c(1,1))
```

Looking at the plots, we can see that our assumptions are met

Intervals for Regression (Answering Question 4)

Prediction interval:

```
predict(stock2.lm,
                newdata = data.frame(Interest_Rate = 2.25,

Unemployment_Rate = 6.01),
                interval = "p")

# the output will look like this:
```

```
#       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.

The prediction interval predicts the response for **one** observation.

Confidence interval:

```
predict(stock2.lm,
                newdata = data.frame(Interest_Rate = 2.25,

Unemployment_Rate = 6.01),
                interval = "c")
# the output will look like this:
#       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.