Multiple Linear Regression

Section 3.2 & 6.1

Dr. Cathy Poliak, cpoliak@uh.edu

University of Houston

Beginning Example

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

- *Interest_Rate*
- *Unemployment_Rate*

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 there a relationship between *stock index price* and *interest rate*?
2. How strong is the relationship between *stock index price* and *interest rate*?
3. Is the relationship linear?
4. How accurately can we predict the *stock index price*?
5. Do both *interest rate* and *unemployment rate* contribute to the *stock index price*?
6. What is the statistical learning problem?

Simple Linear Regression Model

```
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
```

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

Reject Ho, There is evidence that interest rate and stock index price are related.

$\widehat{\text{stock\_index\_price}} = -99.46 + 564.20 * \text{Interest\_Rate}$

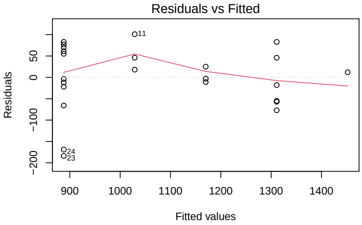
Assumptions about the Model

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

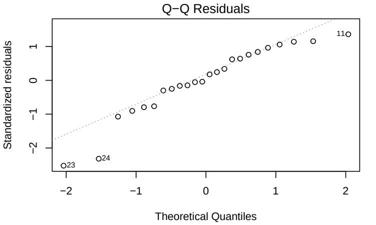
- **L**inear relationship: can we determine a linear relationship between the response an other variables?
- **I**ndependent observations: are the observations a result of a simple random sample?
- **N**ormal distribution: for any fixed value of X, Y is normally distributed.
- **E**qual variance: the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

Plots to Check Assumptions

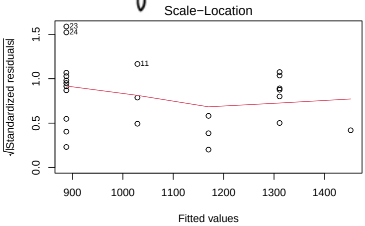
Linear



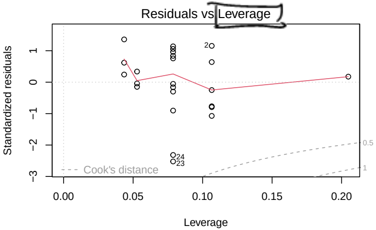
Normal



Equal variance



Extreme value



This does meet the assumption

Using Another Variable

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 Adding the Unemployment Rate and Year

```
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
```

β_0 β_1 β_2 β_3

$\widehat{\text{Stock\_index\_price}} = -56523.71 + 324.59 * \text{Interest\_Rate} - 231.48 * \text{Unemployment Rate} + 28.89 * \text{Year}$

Interpretation of the Estimates

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 $\hat{\beta}_3$.

$\hat{\beta}_3 = 28.89$ For each additional year, the stock index price will increase on average by \$28.89, given Interest rate and unemployment rate are fixed.

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
```

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 accurate 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.
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.

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 $j = 1, 2, \dots p$. That is at least one predictor could be used in the model.

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

F-statistic: $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ vs. H_A: At least one $\beta_j \neq 0$

```
(stock3.f = summary(stock3.lm)$fstatistic)
```

```
 value   numdf   dendf
59.075   3.000  20.000
```

P-value:

```
       value
4.054073e-10
```

Reject Ho, at least one predictor is need in the model.

F-statistic: 59.07 on 3 and 20 DF, p-value: 4.054e-1

Calucation by Hand

1. Residual sum of squares:

$$SSE = RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 103578.7$$

$$= var(stock3.lm\$residual) * (24-1)$$

2. Sum of squares regression:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 917837.1$$

$$= var(stock3.lm\$fitted.values) * (24-1)$$

3. Total sum of squares:

$$SST = TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 1021416$$

$$= var(stock\_index\_price) * (24-1)$$

$$F = \frac{SSR/3}{RSS/(n-p-1)} = \frac{917637.1/3}{103578.7/(24-3-1)} = 59.075$$

$f = 59.075$

$$\text{P-value} = P\left(F_{p,\,n-p-1} \geq f\right)$$

$$= P\left(F_{3,\,20} \geq 59.075\right)$$

$$= 1 - pf(59.075, 3, 20)$$

$$= 4.054\,e^{-10} \approx 0$$

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:

```
                    Estimate   Std. Error    t value   Pr(>|t|)
(Intercept)      -56523.71339 134080.45821 -0.4215656 0.67783849
Interest_Rate       324.58592    123.37418  2.6309065 0.01601745
Unemployment_Rate  -231.48025    127.71974 -1.8124078 0.08496410
Year                 28.89053     66.41674  0.4349887 0.66822831
```

$H_0: \beta_3 = 0$, given that interest rate and unemployment rate is in the model.

$H_A: \beta_3 \neq 0$, give that interest rate and unemployment rate is in the model.

$t = \frac{28.89053 - 0}{66.41674} = 0.435$ P-value $= 2 * P(T \geq t)_{n-p-1} = 2 * P(T \leq -t)$

$= 0.6682$

Fail to reject H_0

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
```

2 Since P-value ≤ 0.1, we reject H_0 and all of the predictors are significant in this model.

1 $\Rightarrow H_0: \beta_1 = \beta_2 = 0$ vs H_A: At least one $\beta_j \neq 0$

Reject H_0, At least one of the predictors is needed in the model.

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).
- 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 (James et al. 2014,P. Bruce and Bruce (2017)):

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).

R Output

```
step(stock3.lm) #stepwise "backwards"
```

①
```
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
```

②
```
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
```

$\hat{y} = 1798.4 + 345.5 * IR - 250.1\ UR$

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

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

- Recall the usual $R^2 = \frac{\overset{\text{SSR}}{\cancel{\text{RSS}}}}{\text{TSS}} = 1 - \frac{\overset{\text{RSS}}{\cancel{\text{SSR}}}}{\text{TSS}}$
- 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.

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	? $\boxed{0.8976}$
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

$$1 - \frac{104559/(24-2-1)}{1021416/23} = 0.8879$$

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.

Calculation of C_p

Given the ANOVA Table:

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

Formula for C_p:

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

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

```r
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
```

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

```
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

The following is an output for the model:

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

```r
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
```

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

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`

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

```
[1] 283.4076
```

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

```
[1] 281.9281
```

```r
BIC(stock3.lm) #Interest_Rate + Unemployement_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