II. The Multiple Regression Model

a. The Multiple Regression Model
b. The Data and Least Squares
c. Inference and F-tests
d. Prediction
e. Multiple Regression Explained: The Pricing Example
f. More on the Interpretation of MR Coefficients
g. Multi-factor Models and Multiple Regression

a. The Multiple Regression Model

Many problems involve more than one independent variable or factor which affects the dependent or response variable

e.g.

- Multiple factor asset pricing models (CAPM vs. APT)
- Demand for a product given prices of competing brands, advertising, household attributes

In the SLR, the conditional mean of Y depends on X. The Multiple Regression Model extends this idea to include more than one independent variable.

a. The Multiple Regression Model

We can add the additional variables into the simple model in a linear fashion:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Conditional Mean: $E[Y|X_1, \ldots, X_k]$

Error Term: Part of Y unrelated to the X's

The interpretation of the regression coefficients, β_j, can be extended from the simple regression case:

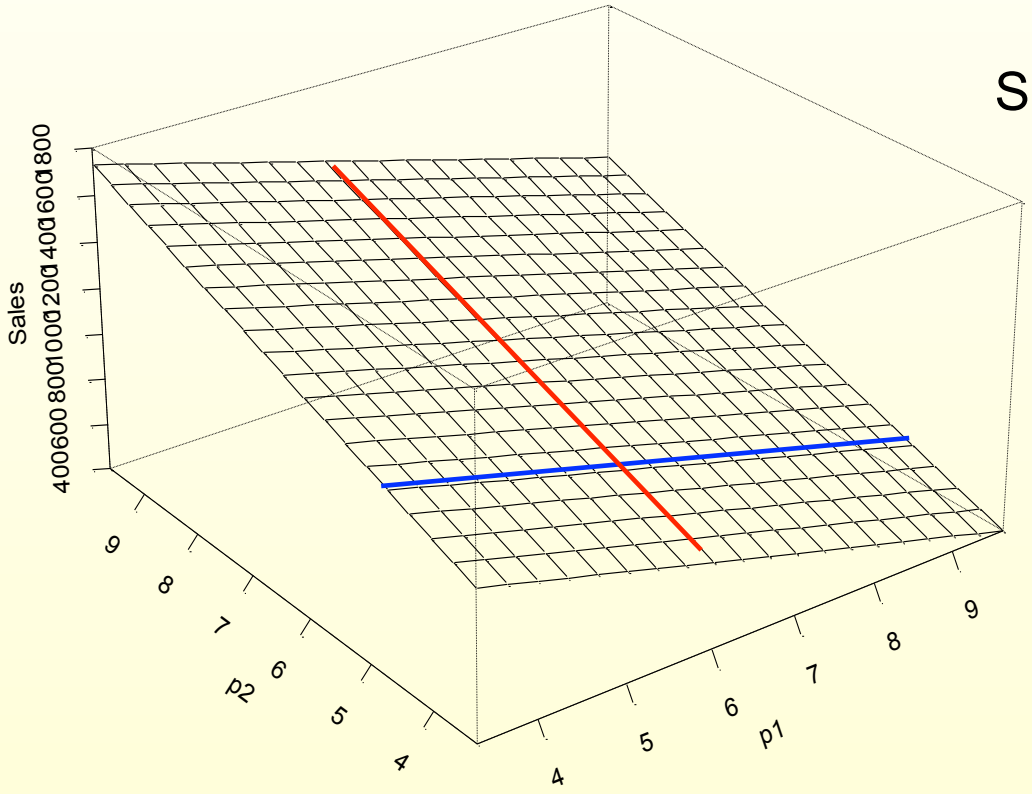
$$\beta_j = \frac{\partial E[Y \mid X_1, \ldots, X_k]}{\partial X_j}$$

Holding all other independent variables constant, the average change in y for a one unit change in x_j is β_j.

a. The Multiple Regression Model

We can plot the regression plane as a surface in three dimensions.

Consider an example of Sales of a product as predicted by Price of this product (P1) and the price of a competing product (P2).



Sales = 1000 – 100P1 + 110P2

hold P2 fixed and vary P1

hold P1 fixed and vary P2

b. The Data and Least Squares

The data in the multiple regression model is a set of points with values of each X variable and Y

Data: $\left(X_{1i}, X_{2i}, \ldots, X_{ji}, \ldots, X_{ki}, Y_i\right) \quad i = 1, \ldots, N$

ith value of X_j in the data

Model:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + \varepsilon_i$$

$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$

b. The Data and Least Squares

How do we estimate the parameters?

We use the principle of Least Squares just as we did in SLR Model:

- we define the fitted values
- find the best fitting plane by minimizing the sum of squared residuals

Fitted Values:

$$\hat{Y}_i = b_0 + b_1 X_{1,i} + \ldots + b_k X_{k,i}$$

Residuals:

$$e_i = Y_i - \hat{Y}_i$$

Least Squares:

$$\text{choose } b_0, b_1, \ldots, b_k \text{ to minimize } \sum_{i=1}^{N} e_i^2$$

Standard Error of the Regression:

$$s = \sqrt{\frac{1}{N-k-1}\sum_{i=1}^{N} e_i^2}$$

b. The Data and Least Squares

Let's run a ***multiple*** regression with the Sales and Price Data.

Model: $\text{Sales}_i = \beta_0 + \beta_1 p1_i + \beta_2 p2_i + \varepsilon_i$

```
> data(multi)
> out=with(multi,
+      lm(Sales~p1+p2)
+ )
> lmSumm(out)
Multiple Regression Analysis:
    3 regressors(including intercept) and 100 observations

lm(formula = Sales ~ p1 + p2)

Coefficients:
            Estimate Std Error t value p value
(Intercept)   115.70     8.548   13.54       0
p1            -97.66     2.669  -36.60       0
p2            108.80     1.409   77.20       0
---
Standard Error of the Regression:  28.42
Multiple R-squared:  0.987  Adjusted R-squared:  0.987
Overall F stat: 3717.29 on 2 and 97 DF, pvalue= 0
```

Note: here I'm using `lmSumm()` function.

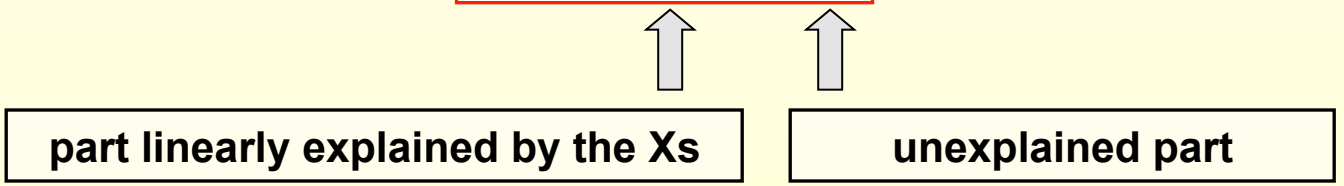
b. The Data and Least Squares

As in the SLR model, the residuals in the multiple regression model are purged of any relationship to the independent variables.

$$\text{corr}\left(X_j, e\right) = 0; \quad j = 1,\ldots,k$$

$$\text{corr}\left(\hat{Y}, e\right) = 0$$

$$Y = \hat{Y} + e$$

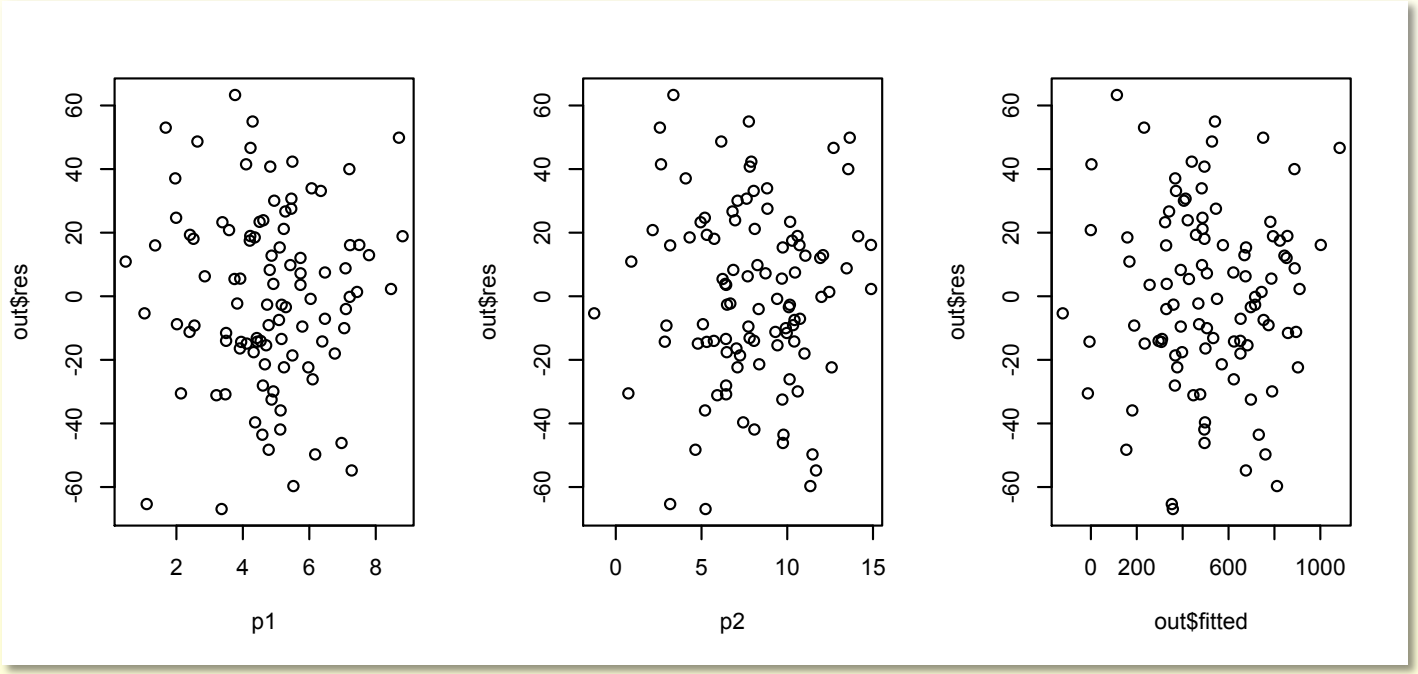


b. The Data and Least Squares

Let's see this graphically:

```
> out=lm(Sales~p1+p2)
> par(mfrow=c(1,3))
> plot(p1,out$res)
> plot(p2,out$res)
> plot(out$fitted,out$res)
```

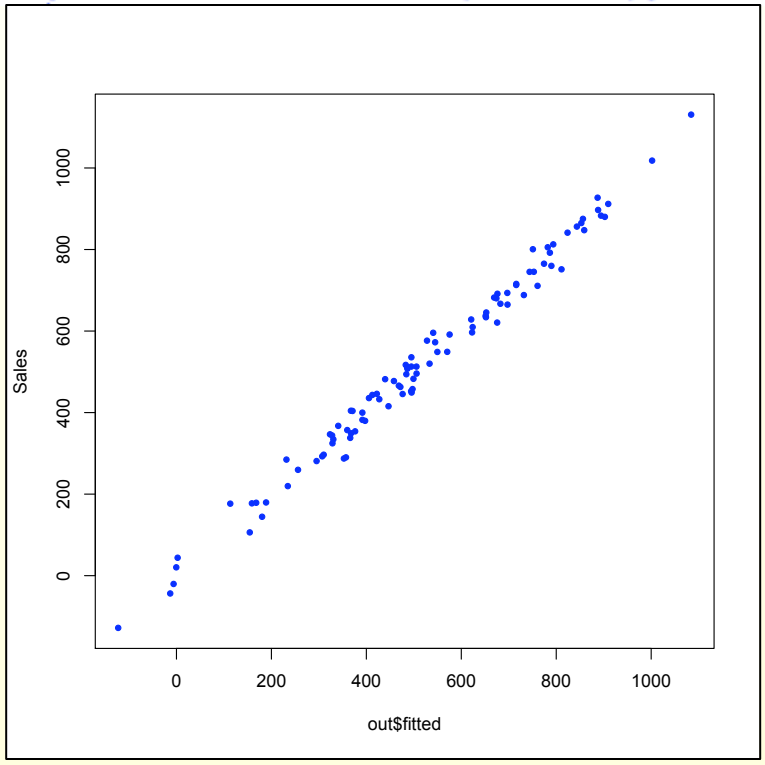
No relationship between the residuals and any of the Xs and also between the fitted and residuals.



b. The Data and Least Squares

If the regression line fits the data well, there should be a strong correlation between the actual and fitted Y values:

```
> plot(out$fitted,Sales,pch=20,col="blue")
```



$$\left(\hat{Y}_i, Y_i\right)$$

```
> corr(multi)
Full Correlation Matrix
          p1   p2 Sales
p1     1.00 0.78  0.44
p2     0.78 1.00  0.90
Sales 0.44 0.90  1.00
```

```
> cor(Sales,out$fitted)
[1] 0.9935395
```

$$R^2 = \left(\text{corr}\left(Y, \hat{Y}\right)\right)^2$$

All inference from the simple regression extends to multiple regression, including:

- **standard errors**
- **confidence intervals**
- **t-tests**

Standard Errors

Since each coefficient estimator is a linear combination of normal random variables, each b_i (i = 0,1, ..., k) is normally distributed.

Notation: $$b_j \sim N(\beta_j, \sigma^2_{b_j}) \quad j = 0, \ldots, k$$

c. Inference and F-tests

The variance of b_j is complicated and depends on the variation of all of the X's.

We call the *estimated* standard deviations the **standard errors**:

$$s_{b_j} = \hat{\sigma}_{b_j}$$

What are the factors which influence the size of the standard errors?

i. σ^2
ii. N
iii. $S_{I.V.}$- variation of X_j which is *independent* of other X vars

(more on independent variation later in these notes)

Confidence Intervals

$$(1-\alpha)\times 100\ \%\text{C.I.:}\quad b_j \pm t_{N-k-1,\alpha/2} s_{b_j}$$

t-Tests

$$\text{test}\quad H_0 : \beta_j = \beta_j^*$$

$$\text{reject}\quad \text{if}\quad t = \left| \frac{b_j - \beta_j^*}{s_{b_j}} \right| \geq t_{N-k-1,\alpha/2}$$

c. F-tests

t-test procedures are designed to examine *one coefficient at a time* for statistical significance.

In many situations, we need a testing procedure that can address *simultaneous* hypotheses about more than one coefficient.

We will look at two important types of simultaneous or joint hypotheses:

1. Overall Test of Significance
2. Partial F test

1. Overall Test of Significance

Suppose we run a regression with some 6 regressors and get a modest R^2.

We may want to see if indeed there is any relationship in this data

Implicitly, we want to test the hypothesis:

$$H_0 : \beta_1 = \beta_2 = \ \ldots = \beta_k = 0$$

c. F-tests: Country Returns Example

Is there any relationship between US equity returns and returns in other countries?

```
lm(formula = usa ~ canada + uk + australia + france + germany +
    japan)

Residuals:
      Min        1Q    Median        3Q       Max
-0.056345 -0.017073  0.000807  0.011979  0.074612

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.006136   0.002309   2.657 0.009171 **
canada       0.444362   0.069587   6.386 5.41e-09 ***
uk           0.225690   0.064915   3.477 0.000753 ***
australia   -0.056688   0.050366  -1.126 0.263061
france       0.166742   0.061338   2.718 0.007733 **
germany     -0.064793   0.057239  -1.132 0.260353
japan       -0.051028   0.034615  -1.474 0.143580
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02257 on 100 degrees of freedom
Multiple R-squared: 0.566,  Adjusted R-squared:  0.54
F-statistic: 21.74 on 6 and 100 DF,  p-value: 3.267e-16
```

Surely the answer must be yes but how strong is evidence?

data(countryret)

c. F-tests

So why don't we just use R^2? If $R^2 > 0$, then we reject the null hypothesis of no relationship.

Problem: R^2 will always be > 0 even if there is no relationship between Y and the X's.

This is because least squares is content to fit "noise" in the data.

This means that SSE will always be less than SST, if only by a small amount.

c. F-tests

To see this, let's generate some garbage data that has nothing to do with USA returns (10 variables):

```
> garbage=matrix(rnorm(107*10,sd=.03,mean=.013),ncol=10)
```

Now we will create a data frame with only the USA return and these 10 variables which are, by definition, independent of USA returns.

```
> summary(lm(usa~.,data=data.frame(cbind(usa,garbage))))
```

`cbind()` takes two spreadsheets and joins them ("column bind").

`data.frame()` takes the spreadsheet (or matrix) and makes it into a data frame.

c. F-tests

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.015368   0.005518   2.785  0.00644 **
V2          -0.157141   0.099389  -1.581  0.11715
V3           0.127047   0.110348   1.151  0.25246
V4          -0.322844   0.110388  -2.925  0.00430 **
V5          -0.152602   0.116177  -1.314  0.19213
V6           0.183102   0.132218   1.385  0.16931
V7           0.017322   0.111435   0.155  0.87680
V8           0.021045   0.109743   0.192  0.84833
V9           0.101637   0.113944   0.892  0.37463
V10          0.054624   0.106672   0.512  0.60978
V11         -0.106342   0.109894  -0.968  0.33564
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 0.03227 on 96 degrees of freedom
Multiple R-squared: 0.1489, Adjusted R-squared: 0.0602
F-statistic: 1.679 on 10 and 96 DF,  p-value: 0.09673
```

Look at that R^2 and t stat

c. F-tests

It seems then, that an R^2 of 15 percent is pretty close to zero in the sense that this can be expected even if the "true" R^2 is 0. As usual, we need a *statistical notion* of how "close is close."

It turns out that under the null hypothesis we can derive the distribution of a transformation of R^2.

Define the **F statistic** for testing overall significance of the regression as follows:

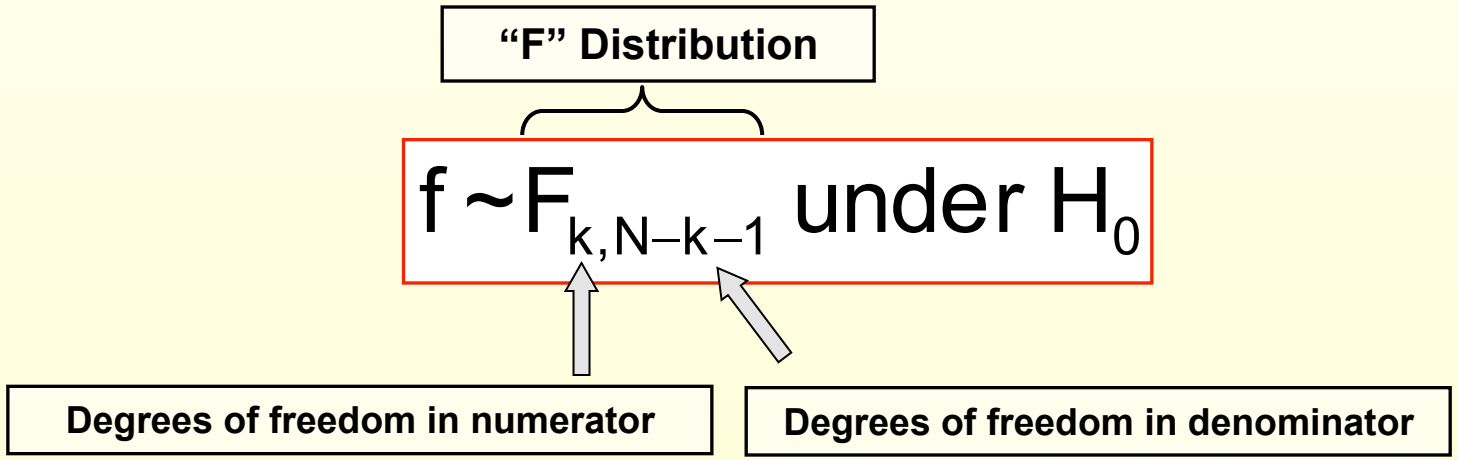
$$f = \frac{R^2 / k}{\left(1 - R^2\right) / \left(N - k - 1\right)}$$

c. F-tests

Properties of this statistic:

- can take on values between 0 and infinity
- the larger R^2 , the larger F
- under the null, we expect F to be clustered around small positive values.

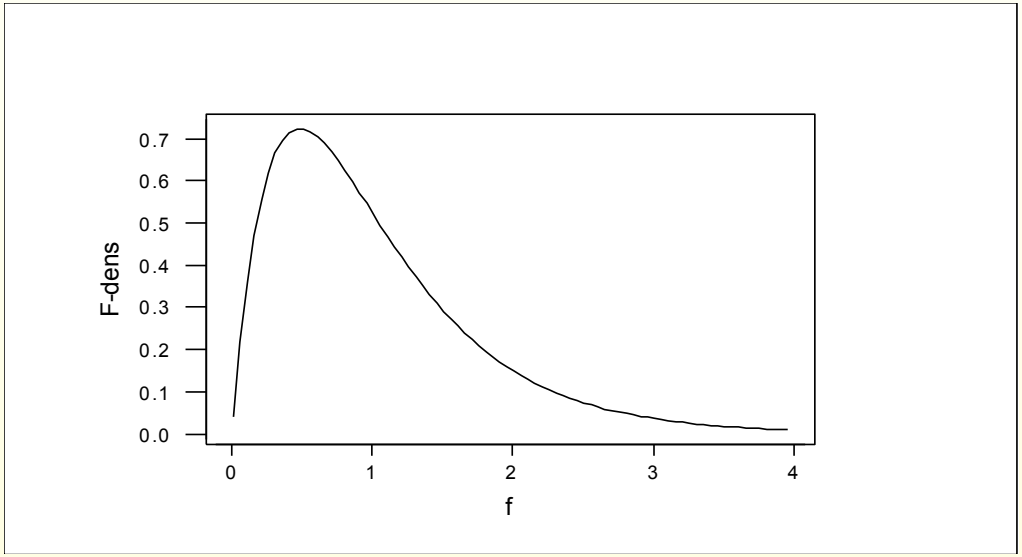
What is the **null distribution of f** introduced on the last slide?



c. F-tests

What kind of distribution is this?

It is a right skewed, positive valued family of distributions which is indexed by two parameters, the df in numerator and the df in the denominator.



To test the null hypothesis, we compute the f statistic using R^2.

We then reject the null if this value of the f statistic is unusually *large.*

c. F-tests

We only get “excited” by large values of F, small values are consistent with the null!.

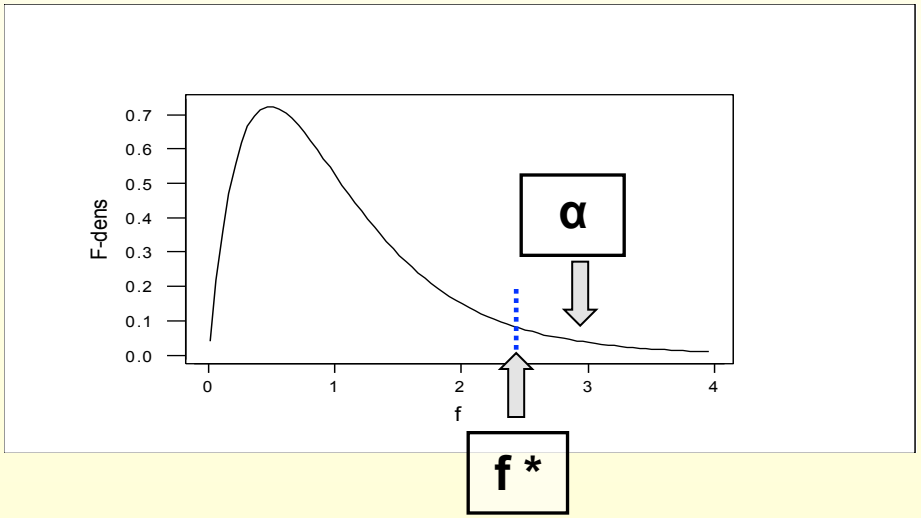
This is called a **one-sided** test.

Summary:

i. choose significance level α

ii. find critical value from INVCDF function…

$$\Pr\left(F_{k,N-k-1} \geq f^{*}_{k,N-k-1,\alpha}\right) = \alpha$$



Summary continued…

iii. Compute f

$$f = \frac{R^2/k}{(1-R^2)/(N-k-1)} = \frac{SSR/k}{SSE/(N-k-1)}$$

Two equivalent expressions for f

Let's do it with country returns data.

Steps i and ii. Set critical value

```
> qf(.95,df1=6,df2=100)
[1] 2.190601
```

c. F-tests

Step iii. Compute F Stat

F is calculated for us by R and is displayed at the bottom of the output. Let's go back to the original regression.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.006136   0.002309   2.657 0.009171 **
canada       0.444362   0.069587   6.386 5.41e-09 ***
uk           0.225690   0.064915   3.477 0.000753 ***
australia   -0.056688   0.050366  -1.126 0.263061
france       0.166742   0.061338   2.718 0.007733 **
germany     -0.064793   0.057239  -1.132 0.260353
japan       -0.051028   0.034615  -1.474 0.143580
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02257 on 100 degrees of freedom
Multiple R-squared: 0.566,  Adjusted R-squared:  0.54
F-statistic: 21.74 on 6 and 100 DF,  p-value: 3.267e-16
```

Is 21.74 large? Look at the p-value which is very close to 0.

2. Partial F – tests

In many situations, a group of regressors are identified which we all agree are important to include in the model.

Another group of variables are somewhat questionable. (these typically have low t-stats)

Are the last group of variables worth including?
We write the model as:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ \dots \ + \beta_{k_1} X_{k_1,i} + \ \beta_{k_1+1} X_{k_1+1,i} + \ \dots \ + \beta_{k_1+k_2} \ X_{k_1+k_2,i} + \varepsilon_i$$

Suspect variables: candidates for deletion

c. F-tests

We wish to test:

$$H_0: \ \beta_{k_1+1} = \ldots = \beta_{k_1+k_2} = 0$$

vs.

$$H_a: \text{at least one of the last } k_2 \beta\text{'s} \neq 0$$

Intuitively, we could check the goodness of fit with and without the k_2 additional variables.

If adding in the additional k_2 variables improved the fit dramatically, we would be tempted to conclude that these variables belong in the model.

c. F-tests

The problem: R^2 will always increase as we add additional variables even if these variables have zero population coefficients.

Why? Least Squares likes additional degrees of freedom to make fitted values which are closer to the observed Y values.

Remember: for the full model, least squares always has the option of shutting down those coefficients.

Adjusted R^2

Some have suggested computing a new quantity called **adjusted R^2** to take into account this problem:

$$\overline{R}^2 = 1 - \frac{SSE/(N-k-1)}{SST/(N-1)} = 1 - \frac{s^2}{s_y^2}$$

Note that we are looking at a ratio of variance estimates. $\overline{R}^2$ will not necessarily increase as other regressors are added.

Also, $\overline{R}^2$ can be < 0!

The problem with the use of $\overline{R}^2$ is that we have no statistical theory on which to base inferences.

Partial F-tests

We can develop a test for inclusion of a subset of variables by using the change in R^2 as we add the variables.

$$f = \frac{\Delta R^2 / k_2}{\left(1 - R^2_{full}\right) / \left(N - \left(k_1 + k_2\right) - 1\right)} \sim F_{k_2, N - k_1 - k_2 - 1}$$

Here: $\Delta R^2 = R^2_{full} - R^2_{restricted}$.

R^2_{full} is the R^2 from the regression with all variables included and $R^2_{restricted}$ is the R^2 from the regression with only the first k_1 variables included

This test is sometimes referred to as a **partial F test** or an **"inclusion/exclusion" test**.

Back to Country Returns Data

1. Run Full and Restricted Regressions

Full regression:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.006136   0.002309   2.657 0.009171 **
canada       0.444362   0.069587   6.386 5.41e-09 ***
uk           0.225690   0.064915   3.477 0.000753 ***
australia   -0.056688   0.050366  -1.126 0.263061
france       0.166742   0.061338   2.718 0.007733 **
germany     -0.064793   0.057239  -1.132 0.260353
japan       -0.051028   0.034615  -1.474 0.143580
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02257 on 100 degrees of freedom
Multiple R-squared: 0.566,  Adjusted R-squared:  0.54
F-statistic: 21.74 on 6 and 100 DF,  p-value: 3.267e-16
```

c. F-tests

Restricted regression (remove insignificant ind vars):

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.006207   0.002297   2.703  0.00805 **
canada      0.410101   0.065075   6.302 7.38e-09 ***
uk          0.163216   0.057742   2.827  0.00565 **
france      0.117555   0.050020   2.350  0.02067 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02273 on 103 degrees of freedom
Multiple R-squared: 0.5469, Adjusted R-squared: 0.5337
F-statistic: 41.44 on 3 and 103 DF,  p-value: < 2.2e-16
```

Compute the Critical Value for F:

N=107 $k_1 = 3$ $k_2 = 3$ Error D.F. = 107-(3+3+1) = 100

```
> qf(.95,df1=3,df2=100)
[1] 2.695534
```

Compute F and P - Values:

$$f = \frac{\Delta R^2 / k_2}{\left(1 - R^2_{full}\right) / \left(N - (k_1 + k_2) - 1\right)} = \frac{(.566 - .547)/3}{(1 - .566)/100} = 1.459$$

```
> pf(1.459,df1=3,df2=100)
[1] 0.769659
```

P value is 1-.770 = .23 Accept Null.

d. Prediction

There is nothing new about prediction in multiple regression that was not covered in the SLR notes.

Prediction Problem:

$$\text{Predict } Y_f \text{ given } X_{1,f}, X_{2,f}, \ldots, X_{k,f}$$
$$\text{and the Data: } \left(Y_i, X_{1,i}, \ldots, X_{k,i}\right) \; i = 1,2,\ldots,N$$

We use the least squares fitted plane to provide the prediction rule. Our prediction interval is:

$$\hat{Y}_f = \left[b_0 + b_1 X_{1,f} + \ldots + b_k X_{k,f}\right] \pm t^*_{N-k-1,\alpha/2} \, s_{pred}$$

This is the standard error of prediction $s_{pred} = \sqrt{\left(s^2 + \text{std}err^2_{fit}\right)}$.

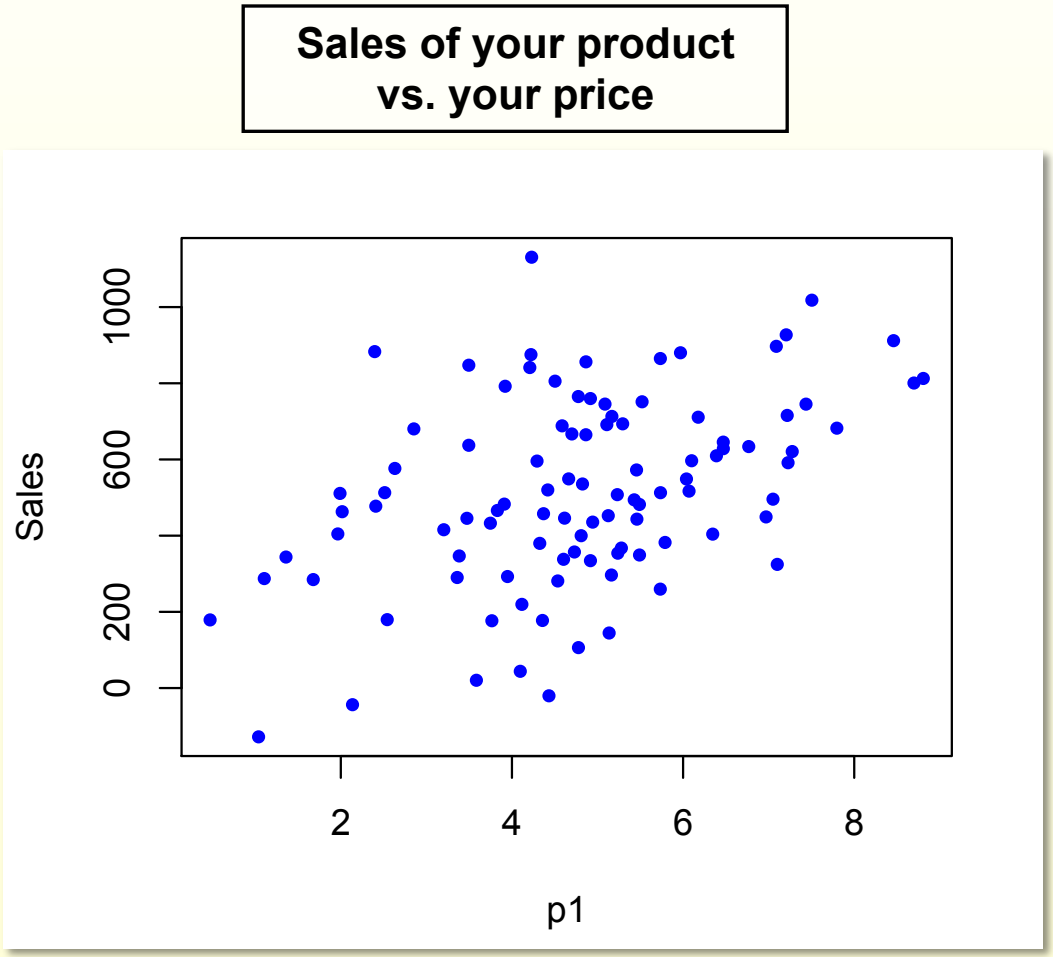
Back to the Pricing and Sales Example:

```
> predict(lm(Sales~p1+p2),new=data.frame(p1=5,p2=8),int="pred")
       fit      lwr      upr
1 497.8298 441.1364 554.5232
```

$$\hat{Y}_f = \left[b_0 + b_1 X_{1,f} + \ldots + b_k X_{k,f} \right] \pm t^*_{N-k-1,\alpha/2} \, s_{pred}$$

e. Multiple Regression Explained: Pricing Puzzle

We've looked only at the multiple regression of Sales on p1 and p2. What if we looked at only the relationship between sales and p1.



e. Multiple Regression Explained: Pricing Puzzle

It appears that there is a **positive,** but weak, relationship between sales and the price of your product.

To verify your visual intuition, you run a regression:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   211.16      66.49   3.176  0.00200 **
p1             63.71      13.04   4.886 4.01e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 223.4 on 98 degrees of freedom
Multiple R-squared: 0.1959, Adjusted R-squared: 0.1877
F-statistic: 23.87 on 1 and 98 DF,  p-value: 4.015e-06
```

e. Multiple Regression Explained

SLR: Sales on p1

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   211.16       66.49   3.176  0.00200
p1             63.71       13.04   4.886 4.01e-06
---
```

Multiple Regression: Sales on p1 and p2

```
Coefficients:
             Estimate Std. Error t value
(Intercept)  115.717       8.548   13.54
p1           -97.657       2.669  -36.59
p2           108.800       1.409   77.20
```

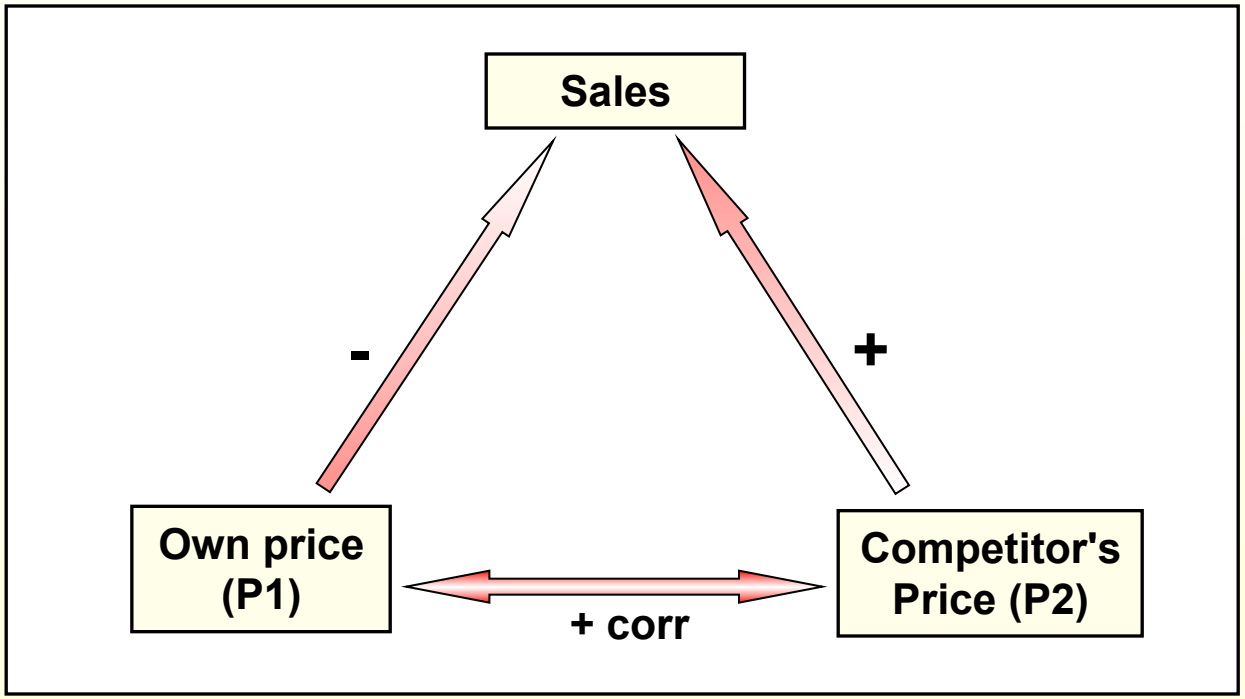
Multiple Reg:

1. neg price effect
2. small std errors

Why is there such a difference?

e. Multiple Regression Explained

Why is there such a difference? The difference stems from confounding of effects.



Multiple regression tries to estimate the **partial** or pure effect of P1 on Sales **controlling** for co-variation with competitor price

e. Multiple Regression Explained

How does the multiple regression work? Let's make a multiple regression using only simple regressions.

If p1 and p2 were uncorrelated, there would be no difference between the simple and multiple regression results.

Problem: How can we "purge" p1 of its relationship to p2?

Solution: Use residuals from regression of p1 on p2

$$p1 = a_0 + a_1 p2 + e_{1.2}$$

This is the part of P1 that is unrelated to P2

Proceed in two steps:

Step 1:

Regress P1 on P2 to purge P1 of relationship to P2.

```
lm(formula = p1 ~ p2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9469 -0.7205  0.1294  0.7971  2.1617

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.49261    0.28628   5.214 1.03e-06 ***
p2           0.41371    0.03316  12.475  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 1.076 on 98 degrees of freedom
Multiple R-squared: 0.6136, Adjusted R-squared: 0.6097
F-statistic: 155.6 on 1 and 98 DF,  p-value: < 2.2e-16

> e_1.2=lm(p1~p2)$residuals
```

Step 2:

Regress Sales on $e_{1.2}$

```
lm(formula = Sales ~ e_1.2)

Residuals:
    Min      1Q  Median      3Q     Max
-638.66 -136.88   11.99  150.49  486.84

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   517.13      22.59  22.893  < 2e-16 ***
e_1.2         -97.66      21.21  -4.604 1.25e-05 ***
```

Same as that from multiple regression

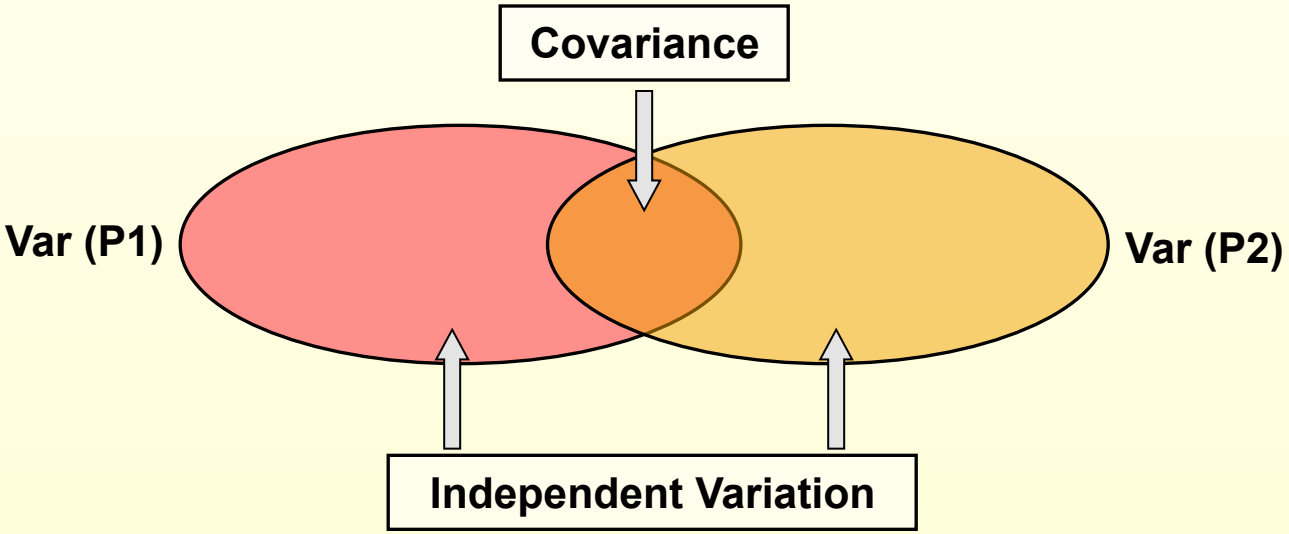
Standard Error is wrong

How Does Multiple Regression Compute the Standard Error?

Key Insight:

The **independent variation** of P1 enables us to estimate the pure or partial effect of P1 on Sales.

Independent variation is that part of the variation in P1 which in unrelated to P2.



f. More on the Interpretation of MR Coefficients

To use MR intelligently, it is essential that we fully understand the sources of the differences between the SLR and MR models.

We also need to understand how to interpret the coefficients for the purpose of making business decisions.

The Difference:

$$MR: \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \ldots \ + \beta_k X_k + \varepsilon$$

$$\beta_1 \neq \beta_1^*$$

$$SLR: \quad Y = \beta_0^* + \beta_1^* X_1 + \varepsilon^*$$

SLR and MR coefficients are only the same if X_1 is uncorrelated with other X vars

f. More on the Interpretation of MR Coefficients

As we saw, the differences between the SLR and MR coefficients stem from the confounding of effects in SLR.

i.e. The SLR model incorporates the influences of all of the other variables in the error term

Thus, we should interpret the SLR coefficients as the effect of X_1 *averaged* over the values and effects of the other variables.

SLR: Effect of X_1 on Y *taking into account the co-movement of* X_1 with the other variables

MR: "Pure" or partial effect of X_1 on Y

f. More on the Interpretation of MR Coefficients

To see this another way, consider a multiple regression with two variables.

Suppose we want to predict Y for a given value of X_1 but we don't have any idea what value X_2 will take on.

A logical way to find the value of X_2 would be to regress X_2 on X_1 and use the fitted value from this regression.

It turns out that if you do this, you will get the same prediction as from the simple regression.

f. More on the Interpretation of MR Coefficients

We use where this predicted value is the expected value of X2 given X1 computed from the simple regression:

$$\hat{X}_2 = \bar{X}_2 + c_1(X_1 - \bar{X}_1)$$

Then we have in the multiple regression:

$$\hat{Y} = \bar{Y} + b_1(X_1 - \bar{X}_1) + b_2(\hat{X}_2 - \bar{X}_2)$$

that can be written as

$$\hat{Y} = \bar{Y} + b_1(X_1 - \bar{X}_1) + b_2c_1(X_1 - \bar{X}_1)$$

or as

$$\hat{Y} = \bar{Y} + (b_1 + b_2c_1)(X_1 - \bar{X}_1)$$

and it can be shown that this equation is identical to the simple linear equation.

f. More on the Interpretation of MR Coefficients

We can demonstrate this in the context of the sales and pricing equation.

```
lm(formula = p2 ~ p1)

Residuals:
     Min       1Q   Median       3Q      Max
-4.59214 -1.36018  0.02994  1.38512  5.54712

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8773     0.6062   1.447    0.151
p1            1.4832     0.1189  12.475   <2e-16
```

And we compute the simple regression coefficient by:

63.7 = -97.7 + 109 (1.48)

b_1: coef on P1 in MR

b_2: coef on P2 in MR

```
The regression equation is

Sales = 116 - 97.7 p1 + 109 p2
```

g. A Multi-factor Model

The CAPM pricing model says that there is only one priced source of risk, i.e. a market factor. That is, expected returns are a function of the beta wrt to the market.

Multi-factor models state that there are multiple (and possibly correlated) *risk factors* which are priced.

There have been many attempts to measure these risk factors.

Let's consider some of the famous Fama-French risk factors.


Some important Risk factors identified in the literature:

SMB: “small” – “big” cap

HML: “Value” – “Growth” based on BE/ME. Value is low BE/ME.

See help file on riskFactors.

```
> corr(riskFactors)
Full Correlation Matrix
      RmRf  SMB  HML
RmRf 1.00 0.33 0.22
SMB  0.33 1.00 0.10
HML  0.22 0.10 1.00
```

Single Factor Model:

```
> lmSumm(outsl)
Multiple Regression Analysis:
    2 regressors(including intercept) and 336 observations

lm(formula = VWNFX ~ RmRf, data = Van_risk)

Coefficients:
            Estimate Std Error t value p value
(Intercept) 0.004216  0.001065    3.96       0
RmRf        0.847100  0.023120   36.64       0
---
Standard Error of the Regression:  0.01934
Multiple R-squared:  0.801  Adjusted R-squared:  0.8
Overall F stat: 1342.49 on 1 and 334 DF, pvalue= 0
```

g. A Multi-factor Model

```
> lmSumm(outml)
Multiple Regression Analysis:
    4 regressors(including intercept) and 336 observations

lm(formula = VWNFX ~ RmRf + SMB + HML, data = Van_risk)

Coefficients:
             Estimate Std Error t value p value
(Intercept)  0.002655 0.0006605    4.02       0
RmRf         0.959000 0.0150100   63.89       0
SMB         -0.193100 0.0216900   -8.90       0
HML          0.425300 0.0232000   18.33       0
---
Standard Error of the Regression:  0.0119
Multiple R-squared:  0.925  Adjusted R-squared:  0.924
Overall F stat: 1365.61 on 3 and 332 DF, pvalue= 0
```

Glossary of Symbols

$F_{\upsilon_1,\upsilon_2}$ - F distribution with u_1 df in the numerator, u_2 df in the denominator

f - value of F statistic

$\bar{R}^2$ - adjusted R-squared

Important Equations

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Multiple Regression Model

$$f = \frac{R^2 / k}{(1 - R^2)/(N - k - 1)} = \frac{SSR/k}{SSE/(N - k - 1)}$$

Overall F-test

$$\overline{R}^2 = 1 - \frac{SSE/(N - k - 1)}{SST/(N - 1)} = 1 - \frac{s^2}{s_y^2}$$

Adjusted R-squared

Important Equations

$$f = \frac{\Delta R^2 / k_2}{\left(1 - R^2_{full}\right) / \left(N - (k_1 + k_2) - 1\right)} \sim F_{k_2, N - k_1 - k_2 - 1} \quad \text{under } H_0$$

Partial or Inclusion/
Exclusion F-test

- `pf(f value, df1=5,df2=54)`: Returns the probability left of value under the F distribution with df of numerator as 5, and df of denominator as 54.
- `qf(prob,df1=5,df2=54)`: Returns the critical value of the probability under the F distribution with df of numerator as 5, and df of denominator as 54.

- `chol2inv(chol(A))`: finds the inverse of the matrix A
- `crossprod(A,B)`: computes A'B efficiently.
- `diag(A)`: fetches diagonal of A
- `A%*%B`: multiplies matrix A by matrix B
- `t(A)`: tranposes the matrix A