MULTIPLE LINEAR REGRESSION

Inference

When we did inference with SLR there was only one slope parameter about which we performed hypothesis testing.

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

For MLR there are multiple slope parameters to investigate.

$$y = \beta_0 + \beta_1 x_1 + +\beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p + \epsilon$$

In the MLR setting with p features, we need to check whether all coefficients are 0.

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_1: $\beta_k \neq 0$ for at least one value of k in 1, 2, ..., p

$$y = \beta_0 + \beta_1 x_1 + +\beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p + \epsilon$$

In the MLR setting with p features, we need to check whether all coefficients are 0.

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_1: $\beta_k \neq 0$ for at least one value of k in 1, 2, ..., p

The null hypothesis case says that there is no useful linear relationship between y and any of the p predictors.

We check with a test based on a statistic that has an F distribution when H_0 is true.

So, we will compute an F statistic.

This is how we will develop the hypothesis test for MLR.

Given $y = \beta_0 + \beta_1 x_1 + +\beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \epsilon$

We will perform three different hypothesis tests for the β_i's.

1] Testing that all of the slope parameters are 0 (F-test)

2] Testing that one slope parameter is 0 (t-test)

3] Testing that a subset (more than one, but not all) of the slope parameters are 0 (F-test)

t-tests for the intercept are rarely relevant

Using inference, we will predict which variables in the model are significant predictors of our response variable.

Three different hypothesis tests for the β_i's.

Does the regression model contain at least one predictor useful in predicting y?

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

H_1: At least one $\beta_j \neq 0$

Is y significantly related to a particular x_i ? (x_1 for instance)

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

Is y significantly linearly related to an x upon controlling for other things?

H_0: $\beta_2 = \beta_3 = 0$

H_1: At least one of $\beta_2, \beta_3 \neq 0$

These are tested with the general linear F-test:

$$F^* = \left(\frac{SSE(R) - SSE(F)}{df_R - df_F}\right) \div \left(\frac{SSE(F)}{df_F}\right)$$

Technically, the second test can also be a t-test

Let's look at some cognitive data for 3-year-olds based on their mother's characteristics (From a study at Duke University). There are 434 rows of data (434 children, $n = 434$).

y	x_1	x_2	x_3	x_4
Score	mom_HS	mom_IQ	mom_work	mom_age

Our MLR line is then of the form: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$

Response variable (dependent): y = Child's score

predictors (independent):

x_1: Did mom attend HS

x_2: Mother's IQ

x_3: Did mother work during the first 3 years of child's life

x_4: Age of mother at child's birth

We will use MLR to answer questions like the following:

1] Is at least one of the features useful in predicting the response?

2] Does each individual feature help to explain the response?
Can we reduce to just a few?

3] How well does the model fit the data?
How well does just a subset of features do?

First test the model as a whole.

Testing that all of the slope parameters are 0.

This is inference for the model as a whole

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ $\quad (y = \beta_0 + \epsilon)$

H_1: At least one β_i is different than 0.

Hypothesis are always about the population parameter(s)

`model.summary` provides…

$F\ statistic = 29.74$

Significance of F, or p-value $= 2.2\times10^{-16}$

Since p-value < 0.05, we reject the null.

The model as a whole is significant;

it contains non-zero β's.

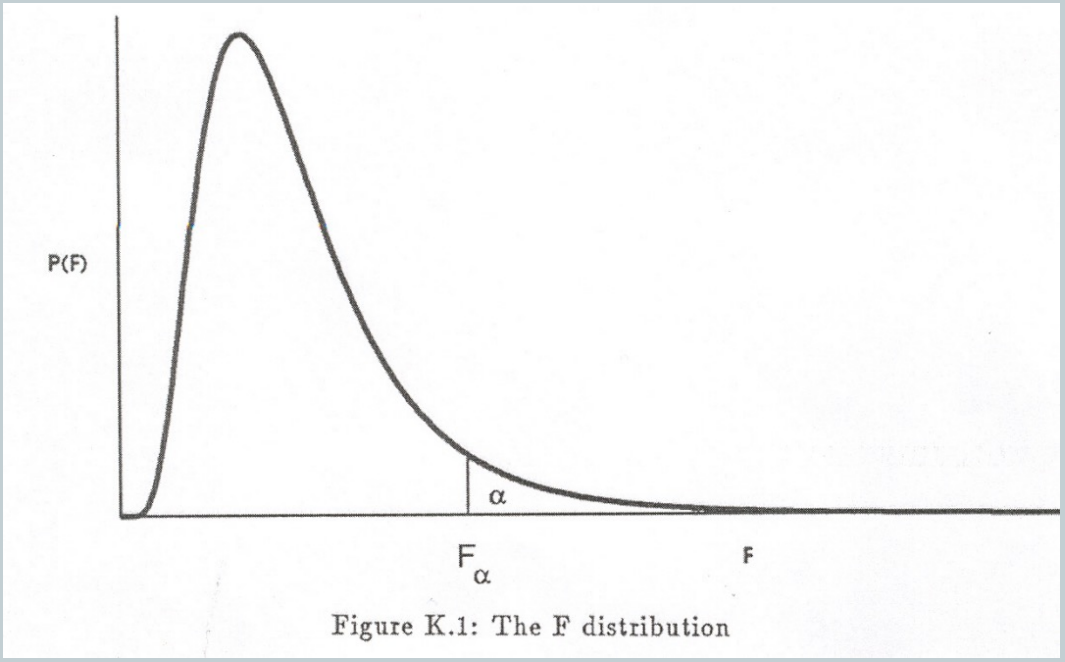


Figure K.1: The F distribution

What is an F-test?

An F-test in regression compares the fits of different linear models. Unlike t-tests that can assess only one regression coefficient at a time, the F-test can assess multiple coefficients simultaneously.

The F-test of the **overall significance** is a specific form of the F-test. It compares a model with no predictors to the model that you specify. A regression model that contains no predictors is also known as an intercept-only model, $y = \beta_0 + \epsilon$.

The hypotheses for the F-test of the overall significance are as follows:

H_0: The fit of the intercept-only model and your model are equal.

H_1: The fit of the intercept-only model is significantly reduced compared to your model.

Why do we use the F-test?

Why compute the p-value for the F-statistic when we could compute p-values for each of the feature slopes?

If we do this, we are testing p different hypotheses instead of a single hypothesis.

At alpha=0.05, how many p-values do we expect to be significant if the null hypothesis is in fact true?

Problem of multiple comparisons!

If the p-value for the F-test of overall significance test is less than your significance level, you can reject the null and conclude that your model provides a better fit than the intercept-only model.

Note, a significant overall F-test could determine that the coefficients are **jointly** not all equal to zero while the tests for **individual** coefficients could determine that all of them are individually equal to zero. Reeses.

Bonus: If the p-value for the overall F-test is less than your significance level, you can conclude that the R^2 value is significantly different from zero.

The F-test

We test the hypothesis with the F-statistic: $$F = \frac{\frac{SST - SSE}{p}}{\frac{SSE}{n-p-1}}$$

Recall:

p is the number of features.

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

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

So, if H_0 is true, then $F \approx 1$.

Whereas if H_1 were true, then we'd have a lower SSE, so SST – SSE is larger, so F is larger. Larger F values are indicative of rejecting the null; i.e. reject all β's being 0.

p-value = `1-stats.f.cdf(F statistic, p features, n-k-1)`

Although, this p-value is provided in the '`.summary`' function.

The F-statistic is a measure of how much better our model is than just using the mean.

Be aware, the F-test yielding a significant result does not mean the model fits the data well, it just means at least one of the β's is non-zero.

The F-test **not** yielding a significant result doesn't mean individual variables included in the model are not good predictors of y, it just means that the combination of these variables doesn't yield a good model.

For the cognitive data the F-test revealed at least one of the β_i's is not 0.

Therefore, there is something worthwhile to look for in this model.

Let's do an individual test on a slope; i.e., any of the β_i's .

We decided the model had merit.

That is, we rejected H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

So, lets check the independent variables:

x_1: Mother attended High School

Is whether or not the mother went to high school a significant predictor of the cognitive test scores of children, given all other variables in the model?

H_0: $\beta_1 = 0$, when all other variables are included in the model.

H_1: $\beta_1 \neq 0$, when all other variables are included in the model.

Estimated $\beta_1 = 5.0948$ (given by `model.summary`)

$t\ value = 2.201$ (this is a t-test)

$std.err = 2.3145$

$p\ value = 0.0282$ → reject the null at $\alpha = 0.05$.

Whether or not mom went to high school **is** a significant predictor of the cognitive test scores of children. Given all other variables in the model.

Individual tests on slopes might look slightly different, but they are the same process as before. The apparent difference between SLR and MLR is in the degrees of freedom.

We use a t-statistic in inference for regression.

$$df = n - k - 1$$

k is the number of predictors.

$$t = \frac{point\ estimate - null\ value}{SE}$$

$$t = \frac{b_1 - 0}{SE_{b_1}}$$

The degrees of freedom is $n - k - 1$, but in SLR it is $n - 2$.

Actually, these are the same.

This is because in SLR $k = 1$.

In other words, in SLR $df = n - k - 1 = n - 1 - 1 = n - 2$.

Along with every predictor for which we calculate a slope estimate we also calculate an intercept and that is where we are losing that one additional degree of freedom.

Start with your sample size, that is the total number of degrees of freedom you have to play with, then you lose the number of DF for the number of predictors you have, then you lose one for the intercept.

1] Hypothesis
2] Significance
3] Sample
4] p-Value
5] Decide

H_0: $\beta_1 = 0$, when all other variables are included in the model.
H_1: $\beta_1 \neq 0$, when all other variables are included in the model.

Estimated $\beta_1 = 5.0948$
$t\ value = 2.20125297$
$std.err = 2.3145$
$p\ value = 0.0282$ → indicates we reject the null.

i.e., $\beta_1 \neq 0$ for x_1 (mom_HS), it is a significant predictor for the child's cognitive score.

$$t = \frac{5.0948 - 0}{2.3145} = 2.20125297$$

$$df = n - k - 1 = 434 - 4 - 1 = 429$$

p-value = `2*(1 - stats.t.cdf(2.20125297, 429))` = 0.028249980284215992

We can also create a confidence interval for slope:

$point\ estimate \pm margin\ of\ error$ aka $b_1 \pm t_{df} \cdot SE_{b_1}$

For instance, suppose the information for mom_work is given as:

Estimated $\beta_3 = 2.53718$ (given by `model.summary`)

$t\ value = 1.079$

$std.err = 2.35067$

$p\ value = 0.2810$ (If this were Hyp. Test, then fail to reject)

We know $df = 434 - 4 - 1 = 429$ so $t_{(.025,429)} = -1.97$

Therefore, our CI is given by $2.54 \pm 1.97 \cdot 2.35 \approx [-2.09, 7.17]$

```
stats.t.ppf(0.025, 429) = -1.965509122975197 8
```

note:

```
stats.norm.ppf(0.025)= -1.9599639845400545
```

Interpreting the 95% CI for the slope of mom_work

Confidence Interval = [-2.09, 7.17] means …

We are 95% confident that, all else being equal, the model predicts that children whose mothers worked during the first three years of their lives score 2.09 points lower to 7.17 points higher than those whose moms did not work.

If this were a hypothesis test,
then we would fail to reject the null and say that it is possible that $\beta_3 = 0$.
Mom working in the first 3 years of life may have no effect on cognitive score.

Where does the MLR information come from?

In Python we will use 'statsmodels' OLS to fit our data with an MLR line.

```
import statsmodels.api as sm
model = sm.OLS(y, X).fit()
model.summary()
```

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.109
Model:	OLS	**Adj. R-squared:**	0.009
Method:	Least Squares	**F-statistic:**	1.091
Date:	Sat, 13 Nov 2021	**Prob (F-statistic):**	0.363
Time:	16:28:52	**Log-Likelihood:**	-494.10
No. Observations:	200	**AIC:**	1030.
Df Residuals:	179	**BIC:**	1099.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.0228	1.669	-1.811	0.072	-6.316	0.271
X1	0.1290	0.193	0.667	0.505	-0.253	0.511
X2	0.8485	0.393	2.161	0.032	0.074	1.623
X3	0.0875	0.362	0.241	0.810	-0.628	0.803
X4	0.2407	0.250	0.964	0.336	-0.252	0.733
X5	-0.0026	0.264	-0.010	0.992	-0.524	0.519
X6	-0.0695	0.199	-0.348	0.728	-0.463	0.324
X7	0.0697	0.382	0.182	0.855	-0.684	0.824
X8	0.3106	0.778	0.399	0.690	-1.225	1.846
X9	-0.1158	0.156	-0.744	0.458	-0.423	0.191
X10	0.1997	0.399	0.501	0.617	-0.587	0.986
X11	-0.2047	0.267	-0.768	0.444	-0.731	0.321
X12	-0.0594	0.246	-0.242	0.809	-0.544	0.426
X13	0.7455	0.780	0.956	0.340	-0.793	2.284

The whole-model summary will look something like this:
(This is not data from cognitive research)

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.109
Model:	OLS	**Adj. R-squared:**	0.009
Method:	Least Squares	**F-statistic:**	1.091
Date:	Sat, 13 Nov 2021	**Prob (F-statistic):**	0.363
Time:	16:28:52	**Log-Likelihood:**	-494.10
No. Observations:	200	**AIC:**	1030.
Df Residuals:	179	**BIC:**	1099.
Df Model:	20		
Covariance Type:	nonrobust		

These are the results for the model as a whole (versus individual results)

F-Statistic should be greater than 1, and the Prob (F-Statistic) is the p-value.

This particular summary indicates that we should not reject the null,

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

F is close to 1, and the Prob(F-Statistic), aka p-value for F, is greater than 0.05.

Notice the summary provides an R^2 and an adjusted R^2.

R-squared:	0.109
Adj. R-squared:	0.009

This is because the R^2 statistic isn't perfect. In fact, it suffers from a major flaw. Its value never decreases no matter the number of variables we add to our regression model.

That is, even if we are adding redundant variables to the data, the value of R^2 does not decrease. It either remains the same or increases with the addition of new independent variables.

This clearly does not make sense because some of the independent variables might not be useful in determining the target variable.

Adjusted R^2 deals with this issue.

The issue/problem:

The standard R^2 value can be artificially inflated by adding lots and lots of frivolous features.

For example, recall the 'color of realtors sign' along with legitimate predictors. And technically, we could fit any set of scatterplot perfectly (overfitting).

Adjusted R^2 takes into account the number of independent variables used for predicting the target variable.

In doing so, we can determine whether adding new variables to the model actually increases the model fit.

The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected.

Typically, the adjusted R-squared is positive, not negative.
It is always lower than the R-squared.

Adding more independent variables or predictors to a regression model tends to increase the R^2 value, which tempts makers of the model to add even more variables.

However, this is called overfitting and can return an unwarranted high R^2 value.

Adjusted R^2 is used to determine how reliable the correlation is and how much it is determined by the addition of independent variables.

Adjusted R^2 is the better model when you compare models that have a different amount of variables.

$$adj.R^2 = 1 - \left[\frac{(1-R^2)(n-1)}{(n-k-1)}\right]$$

n is the number of data points.
k is the number of independent variables.

So, if R^2 does not increase significantly on the addition of a new independent variables, then the value of $adj.R^2$ will actually decrease.

$$adj.R^2 = \left\{1 - \left[\frac{(1-R^2)(n-1)}{(n-\uparrow k-1)\downarrow}\right]\uparrow\right\}\downarrow$$

R^2 not increasing significantly while $adj.R^2$ decreases implies addition of nonsense variables.

On the other hand, if on adding the new independent variable we see a significant increase in R^2 value, then the $adj.R^2$ value will also increase.

$$adj.R^2 = \left\{1 - \left[\frac{\downarrow (1-\uparrow R^2)(n-1)}{(n-k\uparrow -1)\downarrow}\right]\downarrow\right\}\uparrow$$

We can see the difference between R^2 and $adj.R^2$ values if we add a random independent variable to our model.

The addition of meaningful predictors will see both R^2 and $adj.R^2$ increase.

The logic behind it is, that R^2 always increases when the number of variables increases. Meaning that even if you add a useless variable to your model, your R^2 will still increase.

To balance that out, you should always compare models with different number of independent variables with adjusted R^2.

Adjusted R^2 only increases if the new variable improves the model more than would be expected by chance.

Bottom line, if R^2 and adj. R^2 are 'close', then there are no nonsense variables.

Principle of Parsimony:

The principle that the most acceptable explanation of an occurrence, phenomenon, or event is the simplest, involving the fewest entities, assumptions, or changes. K.I.S.

The objective of MLR is not simply to explain the most variation in the data, but to do so with a model with relatively few features that are easily interpreted.

It is thus desirable to adjust R^2 to account for the size of the model, i.e., the number of features.

Quantifying goodness of fit

The coefficient of determination is $R^2 = 1 - \frac{SSE}{SST}$

and it can be adjusted with degrees of freedom for SSE and SST.

$df_{SSE} = n - p - 1$ and $df_{SST} = n - 1$

$$adj.\ R^2 = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}}$$

In both SLR and MLR:

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

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

FYI, other information provided in `.summary`

The AIC is an estimator of prediction error. It estimates the quality of each model.

The BIC is a measure for model selection. Lower BIC are preferred. (A/B Info. Criterion)

Both AIC and BIC attempt to resolve the overfitting problem by introducing a penalty term for the number of parameters in the model.

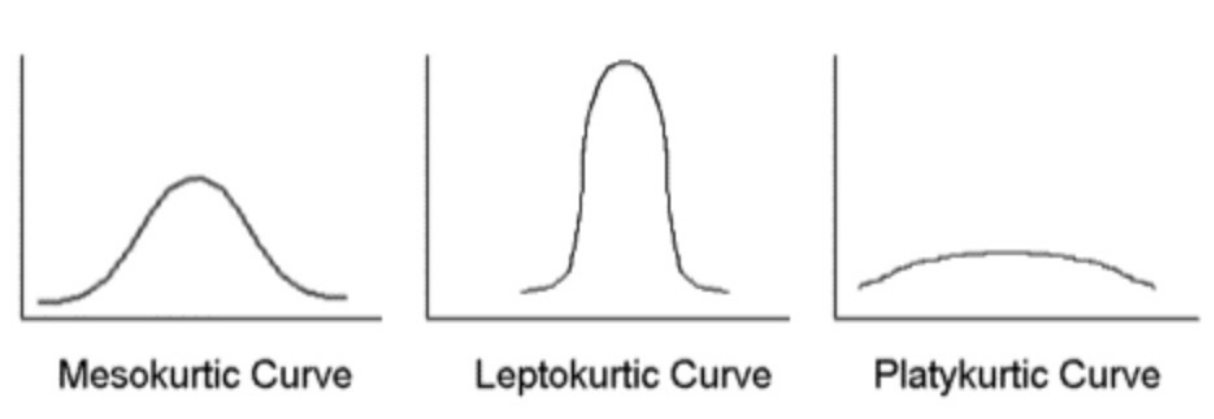
OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.109
Model:	OLS	**Adj. R-squared:**	0.009
Method:	Least Squares	**F-statistic:**	1.091
Date:	Sat, 13 Nov 2021	**Prob (F-statistic):**	0.363
Time:	16:28:52	**Log-Likelihood:**	-494.10
No. Observations:	200	**AIC:**	1030.
Df Residuals:	179	**BIC:**	1099.
Df Model:	20		
Covariance Type:	nonrobust		

FYI, other information provided in `.summary`

Skewness is a measure of symmetry, or lack thereof. Often used for checking the lack of Normality assumption of linear regression. Zero skew is desired. Skewness can cause bad performance in our regression model, i.e., it will affect accuracy.

Kurtosis measures the tails of a distribution relative to the normal. i.e., is the distribution normal, pointy, or flat.



Testing that a subset of the slope parameters are 0

How well does the model fit the data
versus
How well does just a subset of features do?

How do we decide if a reduced model or the full model does a better job of describing the trend in the data ?

A partial F-test is used to determine whether or not there is a statistically significant difference between a regression model and some nested version of the same model.

To determine if two models are significantly different, we perform a partial F-test.

A partial F-test essentially tests whether the group of predictors that you removed from the full model are actually useful and need to be included in the full model.

For the partial F-test:

H_0: All coefficients removed from the full model are zero.

H_1: At least one of the coefficients removed from the full model is non-zero.

If the p-value corresponding to the F test-statistic is below your significance level, then we can reject the null hypothesis and conclude that at least one of the coefficients removed from the full model is significant.

The partial F-test example

Full model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

Reduced model: $y = \beta_0 + \beta_2 x_2 + \beta_4 x_4$

Are the missing features important or are we okay going with the reduced model?

We can answer this with a partial F-test.

H_0: $\beta_1 = \beta_3 = 0$

H_1: at least one of $\beta_1, \beta_3 \neq 0$

Strategy: Fit the full and reduced models.

Determine if the difference in performance is real or just due to chance.

The partial F-test

SSE_{full} = variation unexplained by the full model. p features

$SSE_{reduced}$ = variation unexplained by the reduced model. k features.

Intuitively, if SSE_{full} is much smaller than the $SSE_{reduced}$,
then the full model fits the data much better than the reduced model.

The appropriate test statistic should depend on
the difference $SSE_{reduced} - SSE_{full}$ in unexplained variation.

Test statistic $F = \dfrac{\left(SSE_{reduced} - SSE_{full}\right)/_{(p-k)}}{SSE_{full}/_{(n-p-1)}} \sim F_{p-k,n-p-1}$

Rejection region: $F \geq F_{\alpha, p-k, n-p-1}$

$df_{SSE_{reduced}} - df_{SSE_{full}} =$

$(n-1-k)-(n-p-1) =$

$-k+p =$

$p-k$

Model Selection: Which features should be kept?

Idea: Try all possible combinations of p features and choose the best combo.

But with p features there are 2^p possible models to investigate.

Therefore, for a model with 30 features:

We have a mere $2^{30} = 1{,}073{,}741{,}824$ models to test.

Generally, the variable with the highest correlation is a good predictor.
You can also compare coefficients.
You can look at changes in R^2 value.

Model Selection: Which features should we keep?

Forward Selection: A greedy algorithm for adding features

1] Fit a null model with an intercept but no slopes, $y = \beta_0 + \epsilon$

2] Fit p individual SLR models; 1 for each possible feature.
Add to the null model the one that improves the performance the most, based on some measure. That is, decreases SSE the most, or increases F-statistic the most.

3] Fit p-1 MLR models: 1 for each of the remaining features along with the feature you isolated from step 2.
Add the one that improves model performance the most.

4] Repeat until some stopping criterion is reached.
e.g. some threshold SSE, or some fixed number of features.

Model selection: Which features should we keep?

Backward selection: A greedy algorithm for removing features

1] Fit model with all available features

2] Remove the feature with the largest p-value, i.e. the least significant feature.

3] Repeat until some stopping criterion is reached,
eg. some threshold SSE, or some fixed number of features.

Performing a partial F-test

1] Fit the full regression model and calculate SSE_{full}

2] Fit the nested regression model and calculate $SSE_{reduced}$

3] Perform an ANOVA to compare the full and reduced model.
ANOVA will produce the F test-statistic needed to compare the models.

Next Time: ANOVA

Analysis of Variance

Summary of MLR Testing

There are 3 different hypothesis tests for slopes

1] Testing that all of the slope parameters are 0.
The F-statistic and associated p-value are used for testing whether all of the slope parameters are 0.

2] Testing that one slope parameter is 0.
To test whether one slope parameter is 0, we can use a t-test; look at the p-value.

3] Testing that a subset of the slope parameters are 0.
To test whether a subset of the slope parameters are 0, use the general linear F-test formula by fitting the full model to find SSE(F) and fitting the reduced model to find SSE(R).

The "general linear F-test" involves three basic steps

1] Define a larger full model.
2] Define a smaller reduced model
3] Use an F-statistic to decide whether or not to reject the smaller reduced model in favor of the larger full model.

The null hypothesis always pertains to the reduced model,
while the alternative hypothesis always pertains to the full model.

How do we decide if the reduced model or the full model does a better job of describing the trend in the data when it can't be determined by simply looking at a plot?

We need to quantify how much error remains after fitting each of the two models to our data.

Fit the full model

Obtain the estimates of β_0 and β_1.

Determine the error sum of squares SSE(F)

Fit the reduce model

Obtain the least squares estimate of β_0

Determine the error sum of squares, SSE(R)

Recall that, in general, the error sum squares is obtained by summing the squared distances between the observed and fitted (estimated) responses:

$$\sum(observed\ -fitted)^2$$

Therefore, since y_i is the observed response and $\hat{y}_i$ is the fitted response for the full model:

$$SSE(F) = \sum(y_i - \hat{y}_i)^2$$

And, since y_i is the observed response and $\bar{y}$ is the fitted response for the reduced model:

$$SSE(R) = \sum(y_i - \bar{y})^2$$