L5: Multiple Linear Regression

1. Global ANOVA test.
   - side product: R^2 and adj-R^2
2. t test for individual slope
3. Generalized Linear / Partial F test, focus on
   - Sequential typ = 1.
   - Partial typ = 2.

1. Global ANOVA test:

Summary:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$

H_1: at least one $\beta_i \neq 0$

or

H_0: Reduced model $y_i = \beta_0 + \varepsilon_i$

H_1: Full model: $y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$

Rejection in Global ANOVA test:

$$F_{stat} = \frac{(SSE_R - SSE_F)/(df_R - df_F)}{SSE_F / df_F} \geq F_{\alpha, df_R, df_F}$$

correction: df1= df_sse_reduced- df_sse_full
df2=df_sse_full

indicate substantial reduction in SSE of full model, therefore, "better"

- We are testing if there is any significant predictor in the model.

- we won't know how many or which predictors "contributed" when we have a rejection of H_0.

- More detailed analysis can be seen in the general F tests or t tests.

- ANOVA table for the test

	Sum of squares	d.f.	Mean squares
Regression	$SSR = \Sigma(\hat{y}_i - \bar{y})^2$	$p-1$	$SSR/(p-1)$
Error	$SSE = \Sigma e_i^2$	$n-p$	$SSE/(n-p)$
Total	$SST = \Sigma(y_i - \bar{y})^2$	$n-1$	

$F_{stat} = \frac{MSR}{MSE}$ More general $\frac{(SSE_R - SSE_F)/(df_R - df_F)}{SSE_F/df_F}$

Reject H_0 if $F_{stat} \geq F_\alpha(p-1, n-p)$: there is at least one significant predictor in your model

Interpret the result:

(1) Reject H_0 in this global ANOVA test means we should reject the null model: $y_i = \beta_0 + \varepsilon_i$ and some predictors are significant.

(2) Then the question arises: which predictors are responsible for the rejection of H_0? This can be answered by the initial t-tests of individual coefficients.

II. t test for individual coefficient

Recall: For the OLSE $\vec{b}$ of $\vec{\beta}$, $E(\vec{b}) = \vec{\beta}$, $Var(\vec{b}) = \sigma^2(\vec{X}^T\vec{X})^{-1}$

$\vec{b} = (\vec{X}^T\vec{X})^{-1}\vec{X}^T\vec{y}$ ← linear function of $\vec{y}$

so when σ^2 is known. $\vec{b}$ has a multi-variate normal distribution.

$$\vec{b} \sim MVN(\vec{\beta}, \sigma^2(\vec{X}^T\vec{X})^{-1})$$

When σ^2 is unknown, the variance-covariance matrix of $\vec{b}$ can be estimated by

$$\widehat{Var}(\vec{b}) \overset{or}{=} S^2(\vec{b}) = MSE\,(X^TX)^{-1}$$

Then for each individual $\hat{\beta}_k$, the $var(\hat{\beta}_k)$ can be estimated by $S^2(\hat{\beta}_k) = [MSE\,(X^TX)^{-1}]_{k,k}$

correction: index is (k+1, k+1)

then the studentized t distribution of $\hat{\beta}_k$ is given by

$$\frac{\hat{\beta}_k - \beta_k}{S(\hat{\beta}_k)} \sim t(n-p)$$

d.f. is the same as the d.f. of SSE.

- The t-test and C.I. can be carried out accordingly.
- Reject H_0 when $|t_{stat}| = \left| \frac{\hat{\beta}_k}{\sqrt{MSE\,(X^TX)^{-1}_{k,k}}} \right| \geq t_{\alpha/2,\, n-p}$

Interpret the Result:

(1) The rejection of H_0 indicates X_k is a significant predictor of Y GIVEN the other predictors in the model

(2) It's a partial/marginal test of β_k because $\hat{\beta}_k$ depends on not just X_k, but all the predictors $\vec{X}$.

(3) In a perfect world, when all assumptions are met, the test results are very reliable in the sense that: when we conclude a rejection of H_0, that X_k is a significant predictor, we only have 5% that this a false rejection.

(3) But when assumptions are violated, the test results become much less reliable — Will Discuss in Modeling Problems Later.

summary : R^2 is a side product of the global ANOVA test.

R^2 and adj-R^2

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

where SSE is the SSE of full model.

$SST = \sum (y_i - \bar{y})^2$ doesn't change when fitting different models to the same data.

Fact 1. Recall we have discussed, <u>when adding more predictors,</u> SSE always drops, therefore <u>R^2 always increases.</u>

model 1: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1}$ — only $\sum (x_{i,1} - \bar{x}_1)^2$ contributed the variation

model 2: $\hat{y}_i^* = \hat{\beta}_0^* + \hat{\beta}_1^* x_{i,1} + \hat{\beta}_2^* x_{i,2}$

← more data used

← bigger variation in the model.

← $\sum (x_{i,1} - \bar{x}_1)^2 + \sum (x_{i,2} - \bar{x}_2)^2$

therefore $\sum (y_i - \hat{y}_i)^2 \geq \sum (y_i - \hat{y}_i^*)^2$

Fact 2. R^2 measures the reduction in SSE without considering the lost in dof.

It always suggest a model with more predictors.

Problem: overfitting : adding non-significant predictors that will cause less precise prediction

adding data complexity that impact test results, produces misleading p-values.

adjusted R^2 sometimes yes.

Fact 3. We don't really use R^2 to choose model. but by looking at the value, it helps detecting problems:

R^2 too high: might suggest overfitting.

when R^2 high is normal ? : you are modeling a physical process and have very precise measurements and very routine process, then $R^2 \geq 0.9$ (how the horsepower, engine type impact mpg). is expected.

When R^2 hight is not normal : when the response variable depends on human behavior, predicting it comes with a lot of unexplainable variability. $R^2 \leq 0.5$ is very common.

R^2 too low in the case that it should be higher might suggest:

- underfitting : might add more predictors in the model if available ; might increase sample size if possible.
- Low R^2 is less important than the significance of predictors.

adj-R^2 considers both the reduction in SSE

and loss in degrees of freedom.

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

or

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

← $n-p$ measure the d.f. when p increases.

$$= 1 - \frac{MSE}{MST}$$

Instead of measuring reduction in the sum of squares of error measure the reduction in the mean sum of squares of error.

Suggestion in Model selection:

① adj-R^2 doesn't always increase with number of predictors.

the method "best-subset" of adj-R^2 can give the selection of predictors with highest adj-R^2.

② In linear Regression, model interpretation (significance of predictors) is always more important, we would always look at the test results first before looking at adj-R^2 values.

III. General linear F test to compare reduced model and full model

If you have noticed, in the Python example of Automobile data, where we fit a model

price ~ engine_size + city_mpg + horsepower.

- The model summary table gives the Global F test result:

	SS	df	MS	F	p-value
Regression	*	3	*	251.2	1.03×10^{-16}
Error/Residual	*	195	*		
Total		199-1			

- When we run "sm.stats.anova_lm(reg)"

Instead of giving the above table, it gives:

	df.	SS	MS	F	p-value.
Engine_size	1			724.44	1.3879×10^{-67}
city_mpg	1			23.97	2.0396×10^{-6}
horsepower	1			5.21	2.365×10^{-2}
Residual	195				

When changing types, the value changes!

- The common anova tables in R/Python for MLR are actually comparing models with F test, instead of a global F test on the full model.

For ex. in smf.ols we fit $y \sim X_1 + X_2 + X_3$

and sm.stats.anova_lm(model, typ=1) both have

or

typ=2

similar output.:

	sum of squares	df.	F	p-value.
X_1		1		
X_2		1		
X_3		1		
Residual		$n-4$		

But values are different!

Typ=1: use sequential sum of squares.

X_1: $SS(X_1 \mid \text{Null})$ $H_0: y = \beta_0 + \varepsilon$ $H_1: y = \beta_0 + \beta_1 X_1 + \varepsilon$

X_2: $SS(X_2 \mid X_1)$ $H_0: y = \beta_0 + \beta_1 X_1 + \varepsilon$ $H_1: y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

X_3: $SS(X_3 \mid X_1, X_2)$ $H_0: y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ $H_1: y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$.

The Reject indicate the significance of X_K given the predictors before it are already in the model.

For ex: in anova (typ=1) gives significance of X_2. It means we should choose the model $y \sim X_1 + X_2$ instead of $y \sim X_1$, since adding X_2 would reduce SSE substantially than just having X_1.

Note: 1. all SS are calcated in the above way, therefore, changing the order of the predictors in the ~~the~~ model will CHANGE THE RESULT!

$y \sim X_1 + X_2 + X_3$ $\quad SS(X_2) = SS(X_2 \mid X_1)$

$y \sim X_3 + X_2 + X_1$ $\quad SS(X_2) = SS(X_2 \mid X_3)$

$y \sim X_1 + X_3 + X_2$ $\quad SS(X_2) = SS(X_2 \mid X_1, X_3)$

2. when you change the order, you can see the "importance" of the predictors changes.

3. SSE in the table is the SSE of full model (y ~ all the predictors in the table)

4. sumation of $SS(X_1) + SS(X_2) + SS(X_3) + SSE \neq SST$.

Typ = 2: use partial sum of squares.

X_1 : $SS(X_1 \mid X_2, X_3)$ $H_0: y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ $H_1: y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

X_2 : $SS(X_2 \mid X_1, X_3)$ " $+ \beta_3 X_3 + \varepsilon$

X_3 : $SS(X_3 \mid X_1, X_2)$ "

The Rejection indicate the significance of X_K given all the rest predictors are already in the model.

Note: 1. Doesn't depend on the order of predictors.

2. $F_{stat} = \dfrac{(SSE_R - SSE_F)/1}{MSE_F}$ ← 1 because reduced and full have one predictor difference.

$= t_{stat}^2$

$F_{critical}(\alpha, 1, n-p) = t_{critical}^2(\alpha/2, n-p)$

so p-value from typ=2 is the same as t test for the individual slope.

Summary:

1. Individual t-test test the significance of predictor X_k

using $t_{stat} = \frac{\hat{\beta}_k}{\sqrt{MSE(X^TX)^{-1}_{k,k}}}$ for $H_0: \beta_k = 0$ v.s. $H_1: \beta_k \neq 0$

the result doesn't really consider the importance of other predictors.

2. Global anova test: test if the model fitted is significantly better than null model with only intercept. As long as at least one predictor is significant, this test will be significant.

— Not very informative.

3. Generalized anova with Partial F test.

Typ = 1 (sequential): X_1 F_1

X_2 F_2

X_3 F_3

each F_i gives if adding X_i to $Y \sim X_1 + \ldots + X_{i-1}$ will make it

significantly better.

Hence when changing the order will change the result, since the importance of the predictors to y is different.

Typ = 2 (Not sequential)

		Reduced
X_1	F_1	$y \sim X_2 + X_3$
X_2	F_2	$y \sim X_1 + X_3$
X_3	F_3	$y \sim X_1 + X_2$

each test is comparing the reduced model without the predictor to the model with the predictor.

Note: you already have the sense that with different methods, the feature selection can be complicated. No criteria is the best one. It's more important for us to collect the information along the way. instead just focus on having a predictive model in the end.

"All Models are wrong, but some are useful"
— George Box