P8130: Biostatistical Methods I

Lecture 16: Multiple Linear Regression
ANOVA Testing

Cody Chiuzan, PhD
Department of Biostatistics
Mailman School of Public Health (MSPH)

Recap from Lecture 15

- Multiple Linear Regression
  - Continuous and categorical predictors
  - Interaction terms
- Today's Outline:
  - ANOVA tables and F-tests
  - Testing several coefficients simultaneously
  - Comparing 'nested' models
  - Adjusted and partial R^2

Partitioning the Sum of Squares

- Analysis of variance (ANOVA) vs Regression (practically the same)
- ANOVA results generalize immediately to multiple linear regression

ANOVA table for Simple Linear Regression

Source of Variation	SS	df	MS	$E\{MS\}$
Regression	$SSR = \Sigma(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$\sigma^2 + \beta_1^2 \Sigma(X_i - \bar{X})^2$
Error	$SSE = \Sigma(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	σ^2
Total	$SSTO = \Sigma(Y_i - \bar{Y})^2$	$n-1$		

ANOVA table for MLR

Source of Variation	SS	df	MS	F
Regression	SSR= $\sum_{i-1}^{n}(\hat{Y}_i - \bar{Y})^2$	p	MSR=SSR/p	MSR/MSE
Error	SSE= $\sum_{i-1}^{n}(\hat{Y}_i - \bar{Y})^2$	n-(p+1)	MSE=SSE/(n-p-1)	
Total	SSTO= $\sum_{i-1}^{n}(Y_i - \bar{Y})^2$	n-1		

- *p* represents the number of predictors
- Note: the mean square errors are not additive

ANOVA in SLR

- In SLR, testing that:

$$H_0\colon \beta_1 = 0$$
$$H_1\colon \beta_1 \neq 0$$

- Test statistics: $F^* = MSR/MSE \sim F(1, n-2)$
  - Reject H_0, if $F^* > F(1-\alpha, n-2)$
  - Fail to Reject H_0, if $F^* \leq F(1-\alpha, n-2)$

- In SLR (df=1), the F-test and the t-test are equivalent: $t^2 = F$.

ANOVA in MLR

- Focuses on two types of F-tests.
  - Global (overall) F-test: is there a relationship between Y and the set of covariates:

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

$$H_1\text{: at least one } \beta \text{ is not zero}$$

  - Test statistic: $F^* = \frac{\mathrm{MSR}}{\mathrm{MSE}} > \mathrm{F}(1-\alpha; p, n-p-1)$, reject H_0
  - The null model contains only the intercept:

$$F^* = \frac{\mathrm{MSR}}{\mathrm{MSE}} \leq \mathrm{F}(1-\alpha; p, n-p-1)\text{, fail reject } H_0$$

ANOVA in MLR

- ‘Partial’ F-test; more interesting if you want to test if ‘some’ coefficients are zero at the same time
- Example:

$$H_0\!: \beta_3 = \beta_4 = 0$$

$$H_1\!: \beta_3 \neq 0 \text{ OR } \beta_4 \neq 0$$

- Or even a single coefficient:

$$H_0\!: \beta_3 = 0$$

$$H_1\!: \beta_3 \neq 0$$

ANOVA in MLR

- When using ANOVA for MLR, the order of the terms does matter.

- In ANOVA, the test (p-value) for a certain term is conditioned

  for everything else above.

- In regression, these are conditioned

  for everything else that is in the model.

ANOVA in MLR

- Use ANOVA to test a covariate, adjusting for some, but not other covariates.

- If you want to use ANOVA to test one covariate, adjusting for everything else, then the covariate of interest should be added at the end.

- ‘Partial’ ANOVA is mostly used for nested models, including comparing models with or without interaction terms.

ANOVA in MLR

- 'Partial' F-test for nested models:

$$H_0: small\ model$$

$$H_1: large\ model$$

- Test statistic is given by: $F^* = \frac{(SSR_L - SSR_S)/(df_L - df_S)}{\frac{SSE_L}{df_L}} \sim F_{df_L - df_S, df_L}$, where

$$df_S = n - p_s - 1,\ df_L = n - p_L - 1.$$

- Because $SSTO = SSR + SSE$, the test statistic can also be written as:

$$F^* = \frac{(SSE_S - SSE_L)/(df_L - df_S)}{\frac{SSE_L}{df_L}}$$

- Condition: the two models have to be 'nested', i.e., small model is a 'subset' of the large model.

ANOVA in MLR

- Example: compare Model 1 vs Model 2:

Model 1: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$

Model 2: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$

- We are testing that:

$$H_0\text{: } \beta_3 = \beta_4 = 0$$

$$H_1\text{: } \beta_3 \neq 0 \text{ OR } \beta_4 \neq 0$$

- If we reject the null , we conclude that Model 1 is 'superior'.
- Still, always good to keep in mind the 'principle of parsimony'.

Example ANOVA for MLR

Model 1:

>lm(formula = LOS ~ BEDS + INFRISK, data = data_hosp)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.2703521	0.5038751	12.444	< 2e-16 ***
BEDS	0.0024747	0.0008236	3.005	0.00329 **
INFRISK	0.6323812	0.1184476	5.339	5.08e-07 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.568 on 110 degrees of freedom
Multiple R-squared: 0.3388, Adjusted R-squared: 0.3268
F-statistic: 28.19 on 2 and 110 DF, p-value: 1.31e-10

Model 2:

>lm(formula = LOS ~ BEDS + INFRISK + MS + NURSE, data = data_hosp)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.220159	0.504250	12.335	< 2e-16 ***
BEDS	0.005682	0.001906	2.981	0.00355 **
INFRISK	0.674537	0.118403	5.697	1.07e-07 ***
MS	**0.536804**	**0.509214**	**1.054**	**0.29415**
NURSE	**-0.005905**	**0.002671**	**-2.211**	**0.02918 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.544 on 108 degrees of freedom
Multiple R-squared: 0.3706, Adjusted R-squared: 0.3472
F-statistic: 15.9 on 4 and 108 DF, p-value: 2.924e-10

Example ANOVA for MLR

Comparing Model 1 vs Model 2 (Model 1 is ‘nested’ in Model 2)

R function: anova(reg1, reg2)

Analysis of Variance Table

Model 1: LOS ~ BEDS + INFRISK

Model 2: LOS ~ BEDS + INFRISK + MS + NURSE

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	110	270.56				
2	108	**257.57**	**2**	**12.985**	**2.7223**	**0.07024**

F*=2.72, P=0.07, we fail to reject H0 and conclude that model 2 is not superior, i.e.,

$$\beta_{MS} = \beta_{NURSE} = 0$$

Nested models: Likelihood Ratio Test (LRT)

- When using maximum likelihood estimation, you can also use the LRT.
- The test statistic is given by:

$$\Delta = -2log\frac{L_0}{L_1} = -2(l_0 - l_1) \sim \chi_d^2$$

- In the formula above, *d* represents the difference in the number of parameters between the two models.
- L_0 represents the likelihood for the null model (e.g., small).
- L_1 represents the likelihood for the alternative model (e.g., large).
- l_0 and l_1 represent the log-likelihoods under the null and alternative, respectively.

Nested models: Wald Test

- The Wald test can also be used to test one coefficient or a ‘subset’ of coefficients.
- It is based on the parameter estimates and their standard errors derived from the maximum likelihood or least squares.
- The Wald statistic is inaccurate when the regression coefficient is large, because the standard error tends to be inflated, resulting in the Wald statistic being underestimated.

Nested models: Wald (W) Test

- For a vector of coefficients:

$$H_0\colon \beta = \beta_0$$

- Test statistic is given by:

$$W = (\hat{\beta} - \beta_0)'[var(\hat{\beta})]^{-1}(\hat{\beta} - \beta_0)$$

- Under the null, the test statistic W^* follows a χ^2_p distribution
- In practice, we use an estimate of $var(\hat{\beta})$ and an F distribution
- Note that the Wald and LRT yield similar results, especially when the sample size increases.

Multiple comparison adjustment

Example:

You fit a multiple regression with 20 predictors. By chance alone, you would expect at least one predictor to be 'significant'.

- The main idea is to control the family wise error rate (FWER):

$$FWER = 1 - (1 - \alpha)^k$$

- *k* represents the total number of comparisons.

Multiple comparison adjustment

Solutions for controlling the FWER:

- Use a global (simultaneous) test (F-test)

- Use a Bonferroni adjustment: α/k for each test
  - More conservative -> less powerful

- Define comparisons *a priori* in the statistical plan

- What about using p-values for model selection?
  - Should we remove all non-significant variables? Why/why not?

R^2 vs adjusted R^2

- A high R^2 indicates that the regression model 'fits' the data well.
  - A high percentage in the variance of Y is explained by the relationship with X(s).

- It does not provide information about the model diagnostics.

- Adding a predictor will always increase the R^2
  - To prevent this increase – use adjusted R^2

R^2 vs adjusted R^2

- Unadjusted $R^2 = 1 - \frac{SSE}{SSTO}$

- Adjusted $R^2_{adj} = 1 - \frac{SSE}{SSTO} \cdot \frac{n-1}{n-p-1} = 1 - (1 - R^2) \cdot \frac{n-1}{n-p-1}$

- The adjusted R^2_{adj} increases with a new covariate only if the new term improves the model more than what is expected by chance alone.

Coefficient of partial determination

- 'Partial' R^2 explains the marginal contribution of one predictor X to the variation in Y, when all the other variables are present in the model
- Example for 3 predictors, but it can be generalized for a set of covariates:

$$R^2_{X_1|X_2,X_3} = \frac{SSR(X_1|X_2,X_3)}{SSE(X_2,X_3)}$$

- In R, use the *anova()* function for each of the two models to extract the sums of squares (see Lect 16 R code).