F-tests for Incomplete Data in Multiple Regression Setup

ASHOK CHAURASIA

Advisor: Dr. Ofer Harel
University of Connecticut

OUTLINE

Objective of F-tests in Multiple Linear Regression

In multiple linear regression (MLR), F-tests play a crucial role in testing simultaneous hypotheses. F-tests helps to determine if addition of more predictors has relatively improved the fit.
For example,

- A researcher may be interested testing the association of education and socio-economic variables on suicidal thoughts when the model already contains treatment, race, and gender.
- A nutritionist, who is interested in modeling body fat, may want to test whether mid-arm circumference should be added to a model already containing thigh thickness and triceps skin-fold thickness.
- In an HIV Treatment Adherence program researchers are interested in testing whether alcohol consumption is associated with treatment adherence while accounting for other predictors.

Such research objectives, under usual assumption of MLR with complete data, would be addressed via partial F-tests.

Incomplete Data

Incomplete data are all too common in applied research which complicates the task of testing important research questions.

Methods for handling incomplete data include

- Complete Case Analysis (CCA)
- Single Imputation (Rubin, 1987)
- Weighting Schemes (Rubin, 1987)
- Maximum Likelihood (Little and Rubin, 2002)
- Multiple Imputation (Rubin, 1987)

Model and Hypothesis

For a data, consider the multiple regression model as follows:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where

$\boldsymbol{y}$ is a fully observed n dimensional vector,

$\mathbf{X}$ is the fully observed $n \times p$ matrix of covariates,

$\boldsymbol{\beta} = (\beta_o, \beta_1, \cdots, \beta_k)'$ be a $p = k + 1$ dimensional vector of unknown coefficients, where β_o denotes the intercept, and

$\boldsymbol{\varepsilon} \sim \mathbf{N}_n(0, \sigma^2\mathbf{I})$ with unknown σ^2.

$H_\circ : \beta_{i_1} = \cdots = \beta_{i_r} = 0$ versus H_1: *at least one coefficient is non-zero*,

where

$(i_1, \cdots, i_r) \subseteq (1, \cdots, k)$,

r indicates the number of coefficients, and

$i_1, \cdots, i_r$ denote which coefficient(s) are restricted in $H_\circ$.

THE ISSUE: CONDUCTING PARTIAL F-TESTS WHEN DATA IS INCOMPLETE

Question: *If you have fully observed data then how do you perform simultaneous tests for regression coefficients in MLR*?

Solution: F-tests

Question: *If your data has missing values then how do you handle them*?

Solution: CCA, Single Imputation, Multiple Imputation, MLE, GEE.

Question: *If your data has missing values then how do you perform simultaneous tests for regression coefficients in MLR*?

Solution: ???

F-tests for Fully Observed Data

There are various equivalent ways to define the partial F-test; forms are usually in-terms of regression sums of squares, error sums of squares, or coefficient of determination.

If R_o^2 and R_1^2 represent the coefficient of determination under the null (restricted) and alternative (unrestricted) hypothesis with respective degrees of freedom df_o and df_1, then the partial F-test is defined as

$$F_R = \left(\frac{df_1}{df_{\texttt{restricted}}}\right) \frac{R_1^2 - R_o^2}{1 - R_1^2} \tag{2}$$

where $df_{\texttt{restricted}}$ is the number of parameters restricted in the null hypothesis which is always equal to $df_o - df_1$.
Under the null hypothesis F_R has an F-distribution with degrees of freedom $df_{\texttt{restricted}}$ and df_1.

MULTIPLE IMPUTATION (MI)

Multiple Imputation (Rubin, 1987) is a method for handling missing data in which each missing value is replaced by ($m > 1$) values from the posterior predictive distribution of the missing values given the observed values.

MI comprises of three stages

1. **Imputation**: Multiple Imputed data sets ($\mathscr{D}_1^*, \ldots, \mathscr{D}_m^*$) are created via an Imputation model.
2. **Analysis**: Each imputed data is analyzed (Analysis model) using complete-data techniques for the parameter of interest $\boldsymbol{\theta}$ to yield m point estimates – $Q_1, \ldots, Q_m$, and its variance estimates – $\mathbf{U}_1, \ldots, \mathbf{U}_m$.
3. **Combining**: results from step 2 for each imputed data are combined (Rubin, 1987)

In the scenario where Q is not normally distributed, transformations to approximate normal can be applied, proceeded by combining via Rubin's rules.

COMPUTING R-SQUARE FOR INCOMPLETE DATA VIA MI

Harel (2009) proposed a method to estimate the coefficient of the determination (R^2) from multiply imputed data sets (MIDS).

Suppose $R^2_{o,1}, R^2_{o,2}, \cdots, R^2_{o,m}$ and $R^2_{1,1}, R^2_{1,2}, \cdots, R^2_{1,m}$ denote the coefficient of determination values under $H_\circ$ and H_1, respectively, from the m imputed data sets. Then,

(i) Transform $R^2_{o,j}$ and $R^2_{1,j}$ for $j = 1, \cdots, m$ using Fisher's z-transformation.

$$Q_j = 0.5 \ln \left[\frac{1 + R_j}{1 - R_j} \right] \text{ for } j = 1, \cdots, m \text{ under } H_\circ \text{ and } H_1.$$

(ii) Combine the point and variance estimates using Rubin's rules. Let $\bar{Q}_o$ and $\bar{Q}_1$ denote the combined point estimates under $H_\circ$ and H_1, respectively.

(iii) Back transformation $\bar{Q}_o$ and $\bar{Q}_1$ to obtain coefficient of determination for multiply imputed data under the null (denoted as $\mathfrak{R}^2_o$) and the relative alternative (denoted as $\mathfrak{R}^2_1$).

F-TESTS FOR INCOMPLETE DATA IN MLR SETUP

In light of the estimate the coefficient of determination for incomplete data (via MI) and equation (2), I propose

$$F_{\mathfrak{R}} = \left(\frac{\hat{\nu}_1}{df_\circ - df_1} \right) \frac{\mathfrak{R}_1^2 - \mathfrak{R}_o^2}{1 - \mathfrak{R}_1^2} \tag{3}$$

where

- $\mathfrak{R}_o^2$ and $\mathfrak{R}_1^2$ are estimates for coefficient of determination from MIDS under $H_\circ$ and H_1, respectively, and
- $\hat{\nu}_1$ is the degrees of freedom estimate corresponding to the model under H_1.

Under the null hypothesis $F_{\mathfrak{R}}$ has an approximate F-distribution with numerator degrees of freedom $df_\circ - df_1$ and denominator degrees of freedom $\hat{\nu}_1$. In estimating ν_1, we refer to estimators proposed by Barnard and Rubin (1999) and Reiter (2007).

Simulation Setup

Let $\mathscr{D}_{n,k}$ represent the $n \times k$ data matrix $[\boldsymbol{y}\ \mathbf{X}]$ corresponding to the linear model given by (1) with $k = 3$ predictor variables and sample size $n = 20$ for the hypothesis.

$$H_o : \beta_3 = 0 \ \text{ versus } \ H_1 : \beta_3 \neq 0$$

where β_3 is the regression coefficient of the 3$^{\text{rd}}$ column of $\mathbf{X}$.

1. *Generating Data*: $\mathscr{D}_{n,k} = [\boldsymbol{y}\ \mathbf{X}] \sim \mathbf{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (20.195,\ 25.305,\ 51.170,\ 27.620)'$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\circ$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$ is chosen to reflect data coming from the null ($H_\circ$) and alternative hypothesis (H_1).

$$\boldsymbol{\Sigma}_\circ = \begin{bmatrix} 1 & 0.843 & 0.878 & 0.071 \\ 0.843 & 1 & 0.924 & 0.229 \\ 0.878 & 0.924 & 1 & 0.042 \\ 0.071 & 0.229 & 0.042 & 1 \end{bmatrix} \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.843 & 0.878 & 0.427 \\ 0.843 & 1 & 0.924 & 0.229 \\ 0.878 & 0.924 & 1 & 0.203 \\ 0.427 & 0.229 & 0.203 & 1 \end{bmatrix}$$

$\boldsymbol{\Sigma}_\circ$ and $\boldsymbol{\Sigma}_1$ allow for the assessment of `Type I` (α) and `Type II` (β) errors of our proposed method.

SIMULATION SETUP (CONTINUED...)

2. *Introducing Missingness*: For a given percentage of missingness $(0 < \delta < 1)$, values in $\mathscr{D}_{n,k}$ are made missing at random (MAR; Rubin, 1976) as follows: if u_δ represent the $(1-\delta)100^{th}$ percentile of $\boldsymbol{x}_3$ then,

   (i) one-half of the cases where $x_{3_i} \leq u_\delta$ have $\boldsymbol{y}$ missing, &

   (ii) the remaining cases have $\boldsymbol{x}_1$ missing

   Let $\mathscr{D}_{n,k}^{\text{inc}}$ represent the incomplete data after $\mathscr{D}_{n,k}$ is subjected to the above missingness mechanism.

   Values for percentage of missingness considered are: $\delta = 5,\ 10,\ 15,\ 20,\ 30,\ 40,\ 50.$

SIMULATION SETUP (CONTINUED...)

3. *Imputation*: We assume MAR and multivariate joint normal posterior predictive imputation model based on all the data variables.

4. *Analysis*: For the hypothesis of interest,

   (i) for $\mathscr{D}_{n,k}$, F_{R} is evaluated using (2),

   (ii) for $\mathscr{D}^*_1, \ldots, \mathscr{D}^*_m$, $F_{\mathfrak{R}}$ is evaluated using (3).

5. *Number of Simulations (N)*: Steps 1 to 4 where repeated 1000 times.

SIMULATION RESULTS

δ (%)	Observed Type I error ($\hat{\alpha}$) from $N = 1000$ simulations, using									
	$F_{\mathfrak{R}}$ from MI with								F_R^{CCA}	F_R
	Type	$m = 5$	$m = 25$	$m = 50$	$m = 100$	$m = 250$	$m = 500$	$m = 1000$		
5	BR	0.064*	0.057**	0.063*	0.062*	0.054**	0.061*	0.058*	0.053**	
	Rt	0.196	0.053**	0.057**	0.056**	0.051**	0.056**	0.051**		
10	BR	0.081*	0.065*	0.072*	0.077*	0.078*	0.079*	0.078*	0.046**	
	Rt	0.218	0.057**	0.056**	0.070*	0.070*	0.062*	0.062*		
15	BR	0.052**	0.046**	0.042**	0.037**	0.040**	0.041**	0.040**	0.045**	
	Rt	0.174	0.044**	0.040**	0.033**	0.038**	0.040**	0.039**		
20	BR	0.067*	0.046**	0.044**	0.051**	0.045**	0.046**	0.046**	0.044**	0.057
	Rt	0.178	0.042**	0.038**	0.047**	0.040**	0.043**	0.042**		
30	BR	0.034**	0.047**	0.050**	0.056**	0.057**	0.056**	0.054**	0.047**	
	Rt	0.130	0.040**	0.044**	0.049**	0.052**	0.047**	0.049**		
40	BR	0.024**	0.066*	0.072*	0.069*	0.058*	0.058*	0.063*	0.049**	
	Rt	0.076*	0.053**	0.050**	0.047**	0.041**	0.041**	0.044**		
50	BR	0.062*	0.049**	0.068*	0.047**	0.049**	0.041**	0.049**	0.052**	
	Rt	0.140	0.035**	0.042**	0.028**	0.028**	0.025**	0.026**		

** $\hat{\alpha}$ values $\leq \hat{\alpha}_{F_R}$

* $\hat{\alpha}_{F_R} < \hat{\alpha}$ values ≤ 0.10

Table: Comparison of observed Type I errors of F-statistics corresponding to
(i) fully observed data (F_R),
(ii) complete case analysis (F_R^{CCA}), and
(iii) our proposed MI based method ($F_{\mathfrak{R}}$),
when $H_\circ$ is true for the set-up where $\boldsymbol{y}$ and $\boldsymbol{x}_1$ are missing based on values of $\boldsymbol{x}_3$.

SIMULATION RESULTS (CONTINUED...)

δ (%)	Observed Power $(1-\hat{\beta})$ from $N=1000$ simulations, using									
	$F_{\mathfrak{R}}$ from MI with								F_R^{CCA}	F_R
	Type	$m=5$	$m=25$	$m=50$	$m=100$	$m=250$	$m=500$	$m=1000$		
5	BR	0.631**	0.553**	0.560**	0.557**	0.533**	0.536**	0.560**	0.543**	0.635
	Rt	0.819**	0.541**	0.554**	0.547**	0.524**	0.521**	0.547**		
10	BR	0.438	0.506*	0.545**	0.512**	0.531**	0.537**	0.545**	0.466*	
	Rt	0.637**	0.467*	0.524**	0.491*	0.508**	0.519**	0.529**		
15	BR	0.511**	0.596**	0.545**	0.541**	0.564**	0.565**	0.563**	0.399	
	Rt	0.718**	0.588**	0.532**	0.524**	0.551**	0.559**	0.553**		
20	BR	0.480*	0.563**	0.504*	0.530**	0.560**	0.551**	0.560**	0.337	
	Rt	0.665**	0.550**	0.476*	0.516**	0.545**	0.539**	0.549**		
30	BR	0.474*	0.527**	0.526**	0.530**	0.543**	0.520**	0.536**	0.248	
	Rt	0.682**	0.505*	0.512**	0.511**	0.524**	0.504*	0.515**		
40	BR	0.428	0.567**	0.521**	0.518**	0.526**	0.525**	0.521**	0.183	
	Rt	0.567**	0.549**	0.502*	0.491*	0.495*	0.485*	0.488*		
50	BR	0.507*	0.479*	0.470*	0.450*	0.443	0.423	0.449*	0.139	
	Rt	0.634**	0.425	0.414	0.382	0.372	0.357	0.377		

** Power values $\geq 80\% \left(1-\hat{\beta}_{F_R}\right)$.

* $70\% \left(1-\hat{\beta}_{F_R}\right) \leq$ Power values $< 80\% \left(1-\hat{\beta}_{F_R}\right)$.

Table: Comparison of observed power values of F-statistics corresponding to
(i) fully observed data (F_R),
(ii) complete case analysis (F_R^{CCA}), and
(iii) our proposed MI based method ($F_{\mathfrak{R}}$),
when H_1 is true for the set-up where $\boldsymbol{y}$ and $\boldsymbol{x}_1$ are missing based on values of $\boldsymbol{x}_3$.

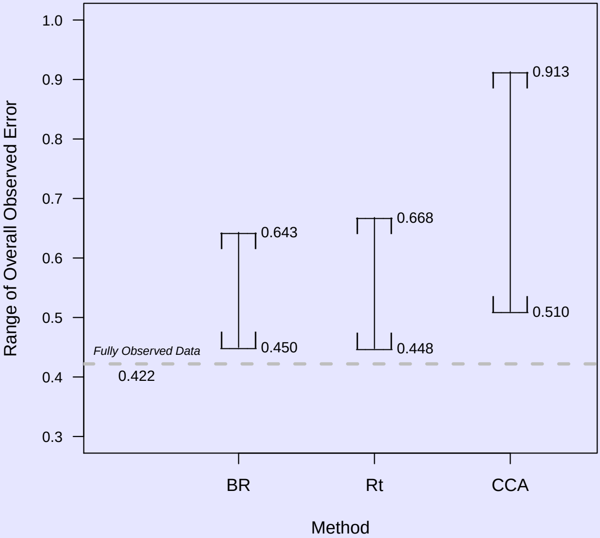

Figure: Comparison of range of observed overall error of the three statistics (from CCA and MI) over all missing percentages when number of imputations $m \geq 25$ and sample size $n = 20$.

CONCLUSION

1. The proposed MI based methods (BR and Rt) have their minimum overall errors (0.45 each) lower than the minimum overall error of CCA (of 0.51), and are much closer to the overall error for fully observed data (0.422).

2. Maximum observed error (`Type I` or `Type II`) with
   - CCA is 91%.
   - MI based method using BR and Rt are 64% and 67%, respectively.
   - CCA at the most is 1.5 times more likely to make wrong decision than our proposed MI based methods.

3. The range for observed overall error
   - is largest for CCA ($0.913 - 0.510 = 0.403$).
   - for MI based method using BR ($0.643 - 0.450 = 0.193$) and Rt ($0.668 - 0.448 = 0.220$) is about one-half of CCA.

Conclusion (continued...)

4. In comparison to the CCA, the probability of making an error is reduced by one-half in our proposed MI based methods.

5. These overwhelmingly positive results of MI based methods correspond to an extreme simulation setting (with small sample size, low power, severe type of missingness) thereby implying that our proposed method's performance will ONLY improve with
   - increasing sample size,
   - increasing distance between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_\circ$,
   - decreasing percentage of missingness, and
   - increasing distance between F-statistic (under null hypothesis) and its corresponding F-critical value.

References

Barnard, J. and D. B. Rubin (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika 86*, 948–955.

Harel, O. (2009). The estimation of $\mathbf{R}^2$ and adjusted $\mathbf{R}^2$ in incomplete data sets using multiple imputation. *Journal of Applied Statistics 36*(10), 1109–1118.

Little, R. and D. Rubin (2002). *Statistical analysis with missing data* (2 ed.). New York: John Wiley & Sons.

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika 94*, 502–508.

Rubin, D. (1976). Inference and missing data. *Biometrika 63*, 581–592.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.