Inference in Multiple Regression: Part 1

- Test for Significant Overall Regression
- Type I Sum of Squares
- Type III Sum of Squares
- Partial F-Test
- Confidence Intervals About Regression Coefficients

Lecture 8
Sections 9.1 – 9.3, 9.5

Three Types of Tests in Multiple Regression

1. **Overall Test:** Does the entire set of independent variables contribute significantly to the prediction of Y?
2. **Test for Addition of a Single Variable:** Does the addition of one particular independent variable add significantly to the prediction of Y after considering all other predictors already in the model?
3. **Test for Addition of a Group of Variables:** Does the addition of some group of independent variables add significantly to the prediction of Y after considering all other predictors already in the model?

Test for Significant Overall Regression

- **Goal:** Determine if the entire set of predictors $X_1, X_2, \dots, X_k$ contributes significantly to the prediction of Y
  - **Procedure:** Test the full model against the model with no predictors where $\bar{Y}$ is the best prediction for all observations
  - **Hypotheses:** $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ vs. H_A: At least one $\beta_i \neq 0$
  - **Test Statistic:** $F = \frac{MSR}{MSE} = \frac{(SSY - SSE)/k}{SSE/(n-k-1)}$
    - Has k and $n-k-1$ df
  - **Equivalent Test Statistic:** $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$

Example: Test for Significant Overall Regression

- **Scenario:** Examine how patient satisfaction (1-100) is related to age (in years), severity of illness (1-100), and anxiety level (1-10)
- **Task:** Perform the test for overall regression.
- **Model:** ________________________________
  - X_1: ____________
  - X_2: ____________
  - X_3: ____________
- **Hypotheses:**
  - H_0: ____________________
  - H_A: ____________________

	C1 Patient Satisfaction	C2 Age	C3 Severity	C4 Anxiety Level
1	48	50	51	7.2
2	57	36	46	7.2
3	66	40	48	6.8
4	70	41	44	5.2
5	89	28	43	5.2
6	36	49	54	9.6
7	46	42	50	6.8
8	54	45	48	7.6
9	26	52	62	9.6
10	77	29	50	6.4

Example: Test for Significant Overall Regression

- **Scenario:** Examine how patient satisfaction (1-100) is related to age (in years), severity of illness (1-100), and anxiety level (1-10)
- **Task:** Perform the test for overall regression.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	9120			0.000
Error	42	4249			
Total	45	13369			

- **Test Statistic:**

$F =$ ____________________________

- **Critical Value:** ______________; **P-Value:** ________
- **Conclusion:** ________ and conclude that the model is ________ ____.

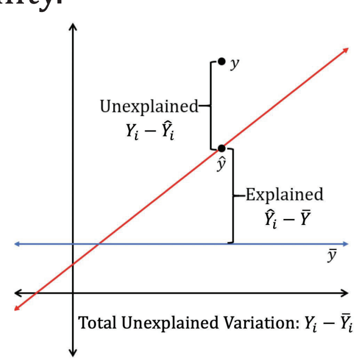
Review: Fundamental Equation of Regression Analysis

- Recall the Fundamental Equation of Regression Analysis:

$$\boldsymbol{SSY = SSR + SSE}$$

where SSR is the amount of variability explained by the regression and SSE is the unexplained variability.

- In multiple regression, every time another predictor is inserted into the model, it helps to explain the response better.
  - **Result:** Some of the ____________________ becomes ____________________



Type I Sum of Squares

- **Type I Sum of Squares:** the amount of explained variability contributed to SSR by a predictor when it is added into a model after considering the contributions to the SSR by the other predictors already added to the model
  - Referred to as **variables added in order** or **sequential** sums of squares
- The types of Type I sums of squares are dependent upon the order in which the variables enter the model. For example:
  - $SS(X_1)$: Sum of squares explained using only X_1
    - X_1 is the first predictor added to the model
  - $SS(X_2|X_1)$: Sum of squares added to SSR by X_2 after X_1 has already been added into the model
  - $SS(X_3|X_1, X_2)$: Sums of squares added to SSR by X_3 after both X_1 and X_2 have been added into the model

Type I Sum of Squares

- Every combination in which predictors are added has its own set of Type I sum of squares
- For a regression with 3 predictors, there are 12 Type I SS:
  - $SS(X_1), SS(X_2), SS(X_3)$
  - $SS(X_1|X_2), SS(X_1|X_3), SS(X_2|X_1), SS(X_2|X_3), SS(X_3|X_1), SS(X_3|X_2)$
  - $SS(X_1|X_2, X_3), SS(X_2|X_1, X_3), SS(X_3|X_1, X_2)$
- Regardless of the order the variables are added, the Type I SS will always sum to the total sum of squares for the regression.
  - $SSR = SS(X_1) + SS(X_2|X_1) + SS(X_3|X_1, X_2)$
- When another predictor is added, its Type I SS comes out of the SSE to guarantee that $SSY = SSR + SSE$ always holds.

Example: Type I Sums of Squares

- **Scenario:** Examine how patient satisfaction (1-100) is related to age (in years), severity of illness (1-100), and anxiety level (1-10)
- **Question:** How can we tell that these are Type I sums of squares?
- **Answer:** Sums of squares for individual predictors ______________
  ____
  - ____
- **Question:** In what order were the predictors added?
- **Answer:**
  1. ________________
  2. ________________
  3. ________________

Source	DF	Seq SS	Seq MS	F-Value	P-Value
Regression	3	9120.5	3040.15	30.05	0.000
Age	1	8275.4	8275.39	81.80	0.000
Severity	1	480.9	480.92	4.75	0.035
Anxiety Level	1	364.2	364.16	3.60	0.065
Error	42	4248.8	101.16		
Total	45	13369.3			

Example: Type I Sums of Squares

- **Scenario:** Examine how patient satisfaction (1-100) is related to age (in years), severity of illness (1-100), and anxiety level (1-10)
- **Question:** What do the Type I sums of squares mean?

Source	DF	Seq SS	Seq MS	F-Value	P-Value
Regression	3	9120.5	3040.15	30.05	0.000
Age	1	8275.4	8275.39	81.80	0.000
Severity	1	480.9	480.92	4.75	0.035
Anxiety Level	1	364.2	364.16	3.60	0.065
Error	42	4248.8	101.16		
Total	45	13369.3			

- **Answer:**
  - **Age:** __________ of the 9120.5 SSR is explained by ____________
  - **Severity:** After ____________________, severity of illness explains an _______________ of the 9120.5 SSR (or ____________________________ SSR)
  - **Anxiety Level:** After accounting for __________________________, anxiety level explains ________ of the 9120.5 SSR (which is __________________)

Example: Type I Sums of Squares

- **Scenario:** Compare the model that include only age and severity of illness with the model that includes all three predictors.
- **Question:** What happened to the sums of squares?

Source	DF	Seq SS	Seq MS	F-Value	P-Value
Regression	2	8756.3	4378.15	40.81	0.000
Age	1	8275.4	8275.39	77.14	0.000
Severity	1	480.9	480.92	4.48	0.040
Error	43	4613.0	107.28		
Total	45	13369.3			

Source	DF	Seq SS	Seq MS	F-Value	P-Value
Regression	3	9120.5	3040.15	30.05	0.000
Age	1	8275.4	8275.39	81.80	0.000
Severity	1	480.9	480.92	4.75	0.035
Anxiety Level	1	364.2	364.16	3.60	0.065
Error	42	4248.8	101.16		
Total	45	13369.3			

- **Answer:** Sum of squares that were _________________________________ _____________ in the model with two predictors got ________________ ____ in the model with all three predictors

Type III Sum of Squares

- **Type III Sum of Squares:** the amount of explained variability contributed to SSR by a predictor assuming it is the last predictor added and all others are already in the model
  - Referred to as **variables added last** or **adjusted** sums of squares
- For a regression with 3 predictors, there are 3 Type III SS:
  - $SS(X_1|X_2, X_3)$: Sum of squares added to SSR by X_1 after X_2 and X_3 are already in the model
  - $SS(X_2|X_1, X_3)$: Sum of squares added to SSR by X_2 after X_1 and X_3 are already in the model
  - $SS(X_3|X_1, X_2)$: Sum of squares added to SSR by X_3 after X_2 and X_3 are already in the model
- <u>*Important*</u>: $SSR \neq SS(X_1|X_2, X_3) + SS(X_2|X_1, X_3) + SS(X_3|X_1, X_2)$

Example: Type III Sum of Squares

- **Scenario:** Examine how patient satisfaction (1-100) is related to age (in years), severity of illness (1-100), and anxiety level (1-10)
- **Question:** How can we tell that these are Type III sums of squares?

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	9120.5	3040.15	30.05	0.000
Age	1	2857.6	2857.55	28.25	0.000
Severity	1	81.7	81.66	0.81	0.374
Anxiety Level	1	364.2	364.16	3.60	0.065
Error	42	4248.8	101.16		
Total	45	13369.3			

- **Answer:** Sums of squares for individual predictors sum to ________
____
  - ______________________________
  - Each adjusted sum of squares assumes the other variables are ________

Example: Type III Sum of Squares

- **Scenario:** Want to investigate how weight is related to height, age, and age squared for children with a nutritional deficiency.
- **Question:** What do the Type III sums of squares mean?

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	9120.5	3040.15	30.05	0.000
Age	1	2857.6	2857.55	28.25	0.000
Severity	1	81.7	81.66	0.81	0.374
Anxiety Level	1	364.2	364.16	3.60	0.065
Error	42	4248.8	101.16		
Total	45	13369.3			

- **Answer:**
  - **Age:** After ______________________________ are inserted into the model, age explains ____________________
  - **Severity:** After ______________________ are inserted into the model, severity of illness explains ____________________
  - **Anxiety Level:** After ________________________ are inserted into the model, anxiety level explains ____________________

Partial F-Test

- **Goal:** Determine if adding a single variable X^* significantly improves the prediction of Y given that $X_1, X_2, \dots, X_p$ are already in the model
  - **Procedure:** Test the full model against the reduced model
    - **Full Model:** Includes $X_1, X_2, \dots, X_p$ as well as X^*
    - **Reduced Model:** Includes only $X_1, X_2, \dots, X_p$ (but not X^*)
  - **Hypotheses:** $H_0: \beta^* = 0$ vs. $H_A: \beta^* \neq 0$
    - *Null Hypothesis:* "X^* does not significantly add to the prediction of Y given that $X_1, X_2, \dots, X_p$ are already predictors in the model so the regression $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + E$ is sufficient."
    - *Alternative Hypothesis:* "X^* significantly adds to the prediction of Y given that $X_1, X_2, \dots, X_p$ are already predictors in the model so the regression $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta^* X^* + E$ is better than the one without X^*."

Partial F-Test

- **Goal:** Determine if adding a single variable X^* significantly improves the prediction of Y given that $X_1, X_2, \dots, X_p$ are already in the model
  - **Sum of Squares:** Regression $SS\left(X^* \middle| X_1, X_2, \dots, X_p\right) =$ Regression $\text{SS}(X_1, X_2, \dots, X_p, X^*) -$ Regression $\text{SS}(X_1, X_2, \dots, X_p)$
    - "Extra sum of squares from adding X^* into the model **equals** regression sum of squares when $X_1, X_2, \dots, X_p$, and X^* are all in the model **minus** regression sum of squares when only $X_1, X_2, \dots, X_p$ are in the model."
  - **Test Statistic:** $F\left(X^* \middle| X_1, X_2, \dots, X_p\right) = \frac{\text{Regression } SS(X^*|X_1, X_2, \dots, X_p)}{MSE(X_1, X_2, \dots, X_p, X^*)}$
    - Has 1 and $n - p - 2$ df
    - "Test statistic **equals** the sum of squares added by X^* given $X_1, X_2, \dots, X_p$ are already in the model **divided by** the mean squared error from the full model that includes $X_1, X_2, \dots, X_p$, and X^*."

Example: Partial F-Test

- **Scenario:** Examine how patient satisfaction (1-100) is related to age (in years), severity of illness (1-100), and anxiety level (1-10)
- **Task:** Test if age (X_1) contributes significantly to the model after severity (X_2) and anxiety level (X_3) have already been included.
- **Models:**
  - **Full Model:** ______________________________
  - **Reduced Model:** ______________________________
- **Hypotheses:** H_0: __________ vs. H_A: __________
- **Need:**
  - ______________________________
  - ______________________________
  - ______________________________

  } $F =$ ____________________

Example: Partial F-Test

- **Outputs with Type I SS:**

Source	DF	Seq SS	Seq MS	F-Value	P-Value
Regression	3	9120.5	3040.15	30.05	0.000
Age	1	8275.4	8275.39	81.80	0.000
Severity	1	480.9	480.92	4.75	0.035
Anxiety Level	1	364.2	364.16	3.60	0.065
Error	42	4248.8	101.16		
Total	45	13369.3			

Source	DF	Seq SS	Seq MS	F-Value	P-Value
Regression	2	6262.9	3131.5	18.95	0.000
Severity	1	4860.3	4860.3	29.41	0.000
Anxiety Level	1	1402.7	1402.7	8.49	0.006
Error	43	7106.4	165.3		
Total	45	13369.3			

- **Test Statistic:** $F =$ ______________________________
- **Critical Value**: ____________________; **P-Value:** __________
- **Conclusion:** ______________ and conclude that ____________________ ____________________ after ______________________________ have been included.

Example: Partial F-Test

- **Scenario:** Examine how patient satisfaction (1-100) is related to age (in years), severity of illness (1-100), and anxiety level (1-10)
- **Task:** Test if severity (X_2) contributes significantly to the model after age (X_1) and anxiety level (X_3) have already been included.
- **Models:**
  - **Full Model:** ______________________________
  - **Reduced Model:** ______________________________
- **Hypotheses:** H_0: ________ vs. H_A: ________
- **Question:** Rather than finding the sums of squares from two different models, how could we find the test statistic?
- **Answer:** Use ______________________________________

Example: Partial F-Test

- **Scenario:** Examine how patient satisfaction (1-100) is related to age (in years), severity of illness (1-100), and anxiety level (1-10)
- **Question:** How can the Type III sum of squares be found for severity of illness?
- **Answer:** Change to ______________________________________
  - $SS(S|A, AL) =$ ______________________________________

Type I (Sequential)

Source	DF	Seq SS	Seq MS
Regression	3	9120.5	3040.15
Age	1	8275.4	8275.39
Severity	1	480.9	480.92
Anxiety Level	1	364.2	364.16
Error	42	4248.8	101.16
Total	45	13369.3	

Source	DF	Seq SS	Seq MS
Regression	2	9038.8	4519.40
Age	1	8275.4	8275.39
Anxiety Level	1	763.4	763.42
Error	43	4330.5	100.71
Total	45	13369.3	

Type III (Adjusted)

Source	DF	Adj SS	Adj MS
Regression	3	9120.5	3040.15
Age	1	2857.6	2857.55
Severity	1	81.7	81.66
Anxiety Level	1	364.2	364.16
Error	42	4248.8	101.16
Total	45	13369.3	

Example: Partial F-Test

- **Scenario:** Examine how patient satisfaction (1-100) is related to age (in years), severity of illness (1-100), and anxiety level (1-10)
- **Task:** Test if severity (X_2) contributes significantly to the model after age (X_1) and anxiety level (X_3) have already been included.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	9120.5	3040.15	30.05	0.000
Age	1	2857.6	2857.55	28.25	0.000
Severity	1	81.7	81.66	0.81	0.374
Anxiety Level	1	364.2	364.16	3.60	0.065
Error	42	4248.8	101.16		
Total	45	13369.3			

- **Test Statistic:** ______________
- **Critical Value:** ______________
- **P-Value:** ____________
- **Conclusion:** ____________ and conclude that severity of illness ______________________________ after age and anxiety level have ________________.

t-Test and Confidence Interval

- **Goal:** Determine if adding a single variable X^* significantly improves the prediction of Y given that $X_1, X_2, \ldots, X_p$ are already in the model
  - **Hypotheses:** $H_0: \beta^* = 0$ vs. $H_0: \beta^* \neq 0$
  - **Test Statistic:** $t = \frac{\hat{\beta}^*}{s_{\hat{\beta}^*}}$ which has $n - p - 1$ df
    - Equivalent to the partial F-test because only one parameter is being tested and $t^2_{n-p-1} = F_{1,n-p-1}$.
  - **Confidence Interval:** A $100(1-\alpha)\%$ confidence interval for the coefficient β^* for the predictor X^* after $X_1, X_2, \ldots, X_p$ have been added is:

$$\hat{\beta}^* \pm t_{n-p-1,1-\alpha/2} \times S_{\hat{\beta}^*}$$

Example: t-Test

- **Task:** Test if anxiety level (X_3) contributes significantly to the model after age (X_1) and severity (X_2) have already been included.

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	151.8	18.3	8.30	0.000	
Age	-1.142	0.215	-5.31	0.000	1.63
Severity	-0.442	0.492	-0.90	0.374	2.00
Anxiety Level	-3.37	1.77	-1.90	0.065	2.01

- **Models:**
  - **Full Model:** ______________________________
  - **Reduced Model:** ______________________________
- **Hypotheses:** H_0: __________ vs. H_A: __________
- **Test Statistic:** ______________________________

Example: t-Test

- **Task:** Test if anxiety level (X_3) contributes significantly to the model after age (X_1) and severity (X_2) have already been included.

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	151.8	18.3	8.30	0.000	
Age	-1.142	0.215	-5.31	0.000	1.63
Severity	-0.442	0.492	-0.90	0.374	2.00
Anxiety Level	-3.37	1.77	-1.90	0.065	2.01

- **Critical Values**: __________________; **P-Value:** __________
- **Confidence Interval:** ______________________________
- **Conclusion:** ______________ and conclude that anxiety level _____ ______________________________ after age and severity of illness have been included.
  - Confidence interval ____________ → ________________