

Measures of Variation in Regression Analysis

MGMT 230: Introductory Statistics

Goals of this section

- Learn in detail how to estimate the relationship between one or more variables.
- Learn how to decompose the variance into variability that is explained and unexplained.

Multiple Regression

- Multiple regression line (population):

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_2 + \dots + \beta_{k-1} x_{k-1} + \epsilon_i$$

- Multiple regression line (sample):

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_2 + \dots + b_k x_k + e_i$$

- k : number of parameters (coefficients) you are estimating.
- ϵ_i : error term, since linear relationship between the x variables and y are not perfect.
- e_i : residual = the difference between the predicted value \hat{y} and the actual value y_i .

Multiple Regression

- Multiple regression line (population):

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_2 + \dots + \beta_{k-1} x_{k-1} + \epsilon_i$$

- Multiple regression line (sample):

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_2 + \dots + b_k x_k + e_i$$

- k : number of parameters (coefficients) you are estimating.
- ϵ_i : error term, since linear relationship between the x variables and y are not perfect.
- e_i : residual = the difference between the predicted value \hat{y} and the actual value y_i .

Multiple Regression

- Multiple regression line (population):

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_2 + \dots + \beta_{k-1} x_{k-1} + \epsilon_i$$

- Multiple regression line (sample):

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_2 + \dots + b_k x_k + e_i$$

- k : number of parameters (coefficients) you are estimating.
- ϵ_i : error term, since linear relationship between the x variables and y are not perfect.
- e_i : residual = the difference between the predicted value \hat{y} and the actual value y_i .

Multiple Regression

- Multiple regression line (population):

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_2 + \dots + \beta_{k-1} x_{k-1} + \epsilon_i$$

- Multiple regression line (sample):

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_2 + \dots + b_k x_k + e_i$$

- k : number of parameters (coefficients) you are estimating.
- ϵ_i : error term, since linear relationship between the x variables and y are not perfect.
- e_i : residual = the difference between the predicted value \hat{y} and the actual value y_i .

Multiple Regression

- Multiple regression line (population):

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_2 + \dots + \beta_{k-1} x_{k-1} + \epsilon_i$$

- Multiple regression line (sample):

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_2 + \dots + b_k x_k + e_i$$

- k : number of parameters (coefficients) you are estimating.
- ϵ_i : error term, since linear relationship between the x variables and y are not perfect.
- e_i : residual = the difference between the predicted value \hat{y} and the actual value y_i .

Multiple Regression

- Multiple regression line (population):

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_2 + \dots + \beta_{k-1} x_{k-1} + \epsilon_i$$

- Multiple regression line (sample):

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_2 + \dots + b_k x_k + e_i$$

- k : number of parameters (coefficients) you are estimating.
- ϵ_i : error term, since linear relationship between the x variables and y are not perfect.
- e_i : residual = the difference between the predicted value \hat{y} and the actual value y_i .

Multiple Regression

- Multiple regression line (population):

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_2 + \dots + \beta_{k-1} x_{k-1} + \epsilon_i$$

- Multiple regression line (sample):

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_2 + \dots + b_k x_k + e_i$$

- k : number of parameters (coefficients) you are estimating.
- ϵ_i : error term, since linear relationship between the x variables and y are not perfect.
- e_i : residual = the difference between the predicted value \hat{y} and the actual value y_i .

Sum of Squares Measures of Variation

- **Sum of Squares Regression (SSR):** measure of the amount of variability in the dependent (Y) variable that is explained by the independent variables (X's).

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

- **Sum of Squares Error (SSE):** measure of the unexplained variability in the dependent variable.

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

Sum of Squares Measures of Variation

- **Sum of Squares Regression (SSR):** measure of the amount of variability in the dependent (Y) variable that is explained by the independent variables (X's).

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

- **Sum of Squares Error (SSE):** measure of the unexplained variability in the dependent variable.

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

Sum of Squares Measures of Variation

- **Sum of Squares Regression (SSR):** measure of the amount of variability in the dependent (Y) variable that is explained by the independent variables (X's).

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

- **Sum of Squares Error (SSE):** measure of the unexplained variability in the dependent variable.

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

Sum of Squares Measures of Variation

- **Sum of Squares Regression (SSR):** measure of the amount of variability in the dependent (Y) variable that is explained by the independent variables (X's).

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

- **Sum of Squares Error (SSE):** measure of the unexplained variability in the dependent variable.

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

Sum of Squares Measures of Variation (continued)

5 / 10

- **Sum of Squares Total (SST):** measure of the total variability in the dependent variable. Does the formula below look familiar?

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

- $SST = SSR + SSE.$

Sum of Squares Measures of Variation (continued)

5 / 10

- **Sum of Squares Total (SST):** measure of the total variability in the dependent variable. Does the formula below look familiar?

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

- SST = SSR + SSE.

Sum of Squares Measures of Variation (continued)

- **Sum of Squares Total (SST):** measure of the total variability in the dependent variable. Does the formula below look familiar?

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

- $SST = SSR + SSE.$

Degrees of Freedom

- Each measure of variability has its own degrees of freedom.
- Degrees of freedom regression = $df_R = k - 1$.
- Degrees of freedom error = $df_E = n - k$.
- Degrees of freedom total = $df_T = n - 1$ (Look familiar?).

Degrees of Freedom

- Each measure of variability has its own degrees of freedom.
- Degrees of freedom regression = $df_R = k - 1$.
- Degrees of freedom error = $df_E = n - k$.
- Degrees of freedom total = $df_T = n - 1$ (Look familiar?).

Degrees of Freedom

- Each measure of variability has its own degrees of freedom.
- Degrees of freedom regression = $df_R = k - 1$.
- Degrees of freedom error = $df_E = n - k$.
- Degrees of freedom total = $df_T = n - 1$ (Look familiar?).

Degrees of Freedom

- Each measure of variability has its own degrees of freedom.
- Degrees of freedom regression = $df_R = k - 1$.
- Degrees of freedom error = $df_E = n - k$.
- Degrees of freedom total = $df_T = n - 1$ (Look familiar?).

Mean Squared Measures of Variation

- **Mean Squared Regression (MSR):** Measure of the *average* amount unexplained variability in the dependent variable. of variability in the dependent (Y) variable that is explained by the independent variables (X's).

$$MSR = \frac{SSR}{df_R}$$

- **Mean Squared Error (MSE):** Measure of the *average* amount of unexplained variability in the dependent variable.

$$MSE = \frac{SSE}{df_E}$$

- No textbook ever talks about a mean squared total, what do you think this would equal?

Mean Squared Measures of Variation

- **Mean Squared Regression (MSR):** Measure of the *average* amount unexplained variability in the dependent variable. of variability in the dependent (Y) variable that is explained by the independent variables (X's).

$$MSR = \frac{SSR}{df_R}$$

- **Mean Squared Error (MSE):** Measure of the *average* amount of unexplained variability in the dependent variable.

$$MSE = \frac{SSE}{df_E}$$

- No textbook ever talks about a mean squared total, what do you think this would equal?

Mean Squared Measures of Variation

- **Mean Squared Regression (MSR):** Measure of the *average* amount unexplained variability in the dependent variable. of variability in the dependent (Y) variable that is explained by the independent variables (X's).

$$MSR = \frac{SSR}{df_R}$$

- **Mean Squared Error (MSE):** Measure of the *average* amount of unexplained variability in the dependent variable.

$$MSE = \frac{SSE}{df_E}$$

- No textbook ever talks about a mean squared total, what do you think this would equal?

Mean Squared Measures of Variation

- **Mean Squared Regression (MSR):** Measure of the *average* amount unexplained variability in the dependent variable. of variability in the dependent (Y) variable that is explained by the independent variables (X's).

$$MSR = \frac{SSR}{df_R}$$

- **Mean Squared Error (MSE):** Measure of the *average* amount of unexplained variability in the dependent variable.

$$MSE = \frac{SSE}{df_E}$$

- No textbook ever talks about a mean squared total, what do you think this would equal?

Mean Squared Measures of Variation

- **Mean Squared Regression (MSR):** Measure of the *average* amount unexplained variability in the dependent variable. of variability in the dependent (Y) variable that is explained by the independent variables (X's).

$$MSR = \frac{SSR}{df_R}$$

- **Mean Squared Error (MSE):** Measure of the *average* amount of unexplained variability in the dependent variable.

$$MSE = \frac{SSE}{df_E}$$

- No textbook ever talks about a mean squared total, what do you think this would equal?

Coefficient of determination

- The **coefficient of determination** is the percentage of variability in y that is explained by x .

$$R^2 = \frac{SSR}{SST}$$

- This *is not the same* as the correlation coefficient.
- In the case of single variable regression, actually equal to the square of the correlation coefficient.
- R^2 will always be between 0 and 1.
- The closer R^2 is to 1, the better x is able to explain y .

Coefficient of determination

- The **coefficient of determination** is the percentage of variability in y that is explained by x .

$$R^2 = \frac{SSR}{SST}$$

- This *is not the same* as the correlation coefficient.
- In the case of single variable regression, actually equal to the square of the correlation coefficient.
- R^2 will always be between 0 and 1.
- The closer R^2 is to 1, the better x is able to explain y .

Coefficient of determination

- The **coefficient of determination** is the percentage of variability in y that is explained by x .

$$R^2 = \frac{SSR}{SST}$$

- This *is not the same* as the correlation coefficient.
- In the case of single variable regression, actually equal to the square of the correlation coefficient.
- R^2 will always be between 0 and 1.
- The closer R^2 is to 1, the better x is able to explain y .

Coefficient of determination

- The **coefficient of determination** is the percentage of variability in y that is explained by x .

$$R^2 = \frac{SSR}{SST}$$

- This *is not the same* as the correlation coefficient.
- In the case of single variable regression, actually equal to the square of the correlation coefficient.
- R^2 will always be between 0 and 1.
- The closer R^2 is to 1, the better x is able to explain y .

Coefficient of determination

- The **coefficient of determination** is the percentage of variability in y that is explained by x .

$$R^2 = \frac{SSR}{SST}$$

- This *is not the same* as the correlation coefficient.
- In the case of single variable regression, actually equal to the square of the correlation coefficient.
- R^2 will always be between 0 and 1.
- The closer R^2 is to 1, the better x is able to explain y .

Coefficient of determination

- The **coefficient of determination** is the percentage of variability in y that is explained by x .

$$R^2 = \frac{SSR}{SST}$$

- This *is not the same* as the correlation coefficient.
- In the case of single variable regression, actually equal to the square of the correlation coefficient.
- R^2 will always be between 0 and 1.
- The closer R^2 is to 1, the better x is able to explain y .

Adjusted R^2

- The more variables you add to the regression, the higher R^2 will be.
- Adding new variables is not necessarily good, when the new variables have nothing to do with the dependent variable.
- The Adjusted R^2 penalizes additional variables.

$$R_{\text{adj}}^2 = 1 - \frac{n-1}{n-k-1} (1 - R^2)$$

- When the adjusted R^2 increases when adding a variable, then the additional variable really did help explain the dependent variable.
- When the adjusted R^2 decreases when adding a variable, then the additional variable does not help explain the dependent variable.

Adjusted R^2

- The more variables you add to the regression, the higher R^2 will be.
- Adding new variables is not necessarily good, when the new variables have nothing to do with the dependent variable.
- The Adjusted R^2 penalizes additional variables.

$$R_{\text{adj}}^2 = 1 - \frac{n-1}{n-k-1} (1 - R^2)$$

- When the adjusted R^2 increases when adding a variable, then the additional variable really did help explain the dependent variable.
- When the adjusted R^2 decreases when adding a variable, then the additional variable does not help explain the dependent variable.

Adjusted R^2

- The more variables you add to the regression, the higher R^2 will be.
- Adding new variables is not necessarily good, when the new variables have nothing to do with the dependent variable.
- The Adjusted R^2 penalizes additional variables.

$$R_{\text{adj}}^2 = 1 - \frac{n-1}{n-k-1} (1 - R^2)$$

- When the adjusted R^2 increases when adding a variable, then the additional variable really did help explain the dependent variable.
- When the adjusted R^2 decreases when adding a variable, then the additional variable does not help explain the dependent variable.

Adjusted R^2

- The more variables you add to the regression, the higher R^2 will be.
- Adding new variables is not necessarily good, when the new variables have nothing to do with the dependent variable.
- The Adjusted R^2 penalizes additional variables.

$$R_{\text{adj}}^2 = 1 - \frac{n-1}{n-k-1} (1 - R^2)$$

- When the adjusted R^2 increases when adding a variable, then the additional variable really did help explain the dependent variable.
- When the adjusted R^2 decreases when adding a variable, then the additional variable does not help explain the dependent variable.

Adjusted R^2

- The more variables you add to the regression, the higher R^2 will be.
- Adding new variables is not necessarily good, when the new variables have nothing to do with the dependent variable.
- The Adjusted R^2 penalizes additional variables.

$$R_{\text{adj}}^2 = 1 - \frac{n-1}{n-k-1} (1 - R^2)$$

- When the adjusted R^2 increases when adding a variable, then the additional variable really did help explain the dependent variable.
- When the adjusted R^2 decreases when adding a variable, then the additional variable does not help explain the dependent variable.

Adjusted R^2

- The more variables you add to the regression, the higher R^2 will be.
- Adding new variables is not necessarily good, when the new variables have nothing to do with the dependent variable.
- The Adjusted R^2 penalizes additional variables.

$$R_{\text{adj}}^2 = 1 - \frac{n-1}{n-k-1} (1 - R^2)$$

- When the adjusted R^2 increases when adding a variable, then the additional variable really did help explain the dependent variable.
- When the adjusted R^2 decreases when adding a variable, then the additional variable does not help explain the dependent variable.

- Problems computing regression:
 - Section 13.2, page 522: problems 13.4 through 13.7.
- Computing Coefficient of determination.
 - Section 13.3, pages 528-529, problems 13.16 through 13.19.