STAT 5000

Statistical Methods I

Week 12

Fall 2024

Dr. Danica Ommen

Unit 3

Multiple Linear Regression

Introduction

Notation

- $i = 1, \ldots, n$: number of observations.
- Y_i: quantitative response variable
- $x_{i1}, x_{i2}, \ldots, x_{ik}$: k explanatory variables
- Values of $x_{i1}, x_{i2}, \ldots, x_{ik}$ are treated as known and fixed

Research Questions

- Does the MLR model significantly explain the response variable Y_i and how well does it explain the variation in the response variable Y_i?
- Which explanatory variables are significant in the MLR model?
- Which set of explanatory variables are significant in the MLR model?
- What value of the conditional mean of Y_i would we predict for given values of $x_{i1}, x_{i2}, \ldots, x_{ik}$?
- What value of Y_i would we predict for given values of $x_{i1}, x_{i2}, \ldots, x_{ik}$?

MLR Model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

MLR Assumptions

- Fixed values of the explanatory variables, $x_{i1},\ x_{i2}, \cdots x_{ik}$
- Conditional mean of Y given the values of $x_{i1}, x_{i2}, \ldots, x_{ik}$ is linear: $\mu_{Y|x_{i1},x_{i2},\ldots,x_{ik}} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$
- Additive random errors: $Y_i = \mu_{Y|x_{i1},x_{i2},\ldots,x_{ik}} + \epsilon_i$
- Independent (uncorrelated) random errors
- Homogeneous error variance: $Var(\epsilon_i) = \sigma^2$
- Normally distributed random errors: $\epsilon_i \sim N(0, \sigma^2)$

Assumptions

- Conditional distribution of Y_i for a given set of values $x_{i1}, x_{i2}, \ldots, x_{ik}$ is

$$N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_k x_{ik}, \sigma^2)$$

- Equivalently, we have $\mathbf{Y} \sim MVN(X\boldsymbol{\beta}, \sigma^2 I_n)$.

Parameters (Coefficients)

- $\beta_j =$ population slope for explanatory variable x_j
  - Change in the conditional mean of Y for a one unit increase in x_j, *holding all other explanatory variables constant*
  - Linear effect of x_j on conditional mean of Y *after adjusting for linear effect of the other predictors on Y and linear effects of the other explanatory variables on x_j.*
- $\beta_0 =$ population intercept
  - the conditional mean of Y when $x_1 = x_2 = \cdots = x_k = 0$

Parameters (Coefficients)

- Interpretation of parameters $\beta_0, \beta_1, \ldots, \beta_k$ depends on the presence or absence of other explanatory variables in the model
- Example:
  - Model 1: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_k x_{ik} + \epsilon_i$
  - Model 2: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$
- Interpretation of parameters β_0, β_1, and β_2 are NOT the same in the two models

Parameters (Variance)

- σ^2 is the variation of responses about the conditional mean of Y for any specific values of $x_1, x_2, \ldots, x_k$

Least Squares Estimation

Find **b**, the least squares estimator for β, that minimizes

$$\begin{aligned} q(\mathbf{b}) &= \sum_{i=1}^{n}(Y_i - b_o - b_1 x_{i1} - \cdots - b_k x_{ik})^2 \\ &= (\mathbf{Y} - X\mathbf{b})^T(\mathbf{Y} - X\mathbf{b}) = \mathbf{e}^T\mathbf{e} \end{aligned}$$

where $\mathbf{e} = \mathbf{Y} - X\mathbf{b}$ is the vector of residuals

- Solve the set of normal equations: $(X^T X)\mathbf{b} = X^T\mathbf{Y}$
- Solution: assuming X is of full column rank

$$\mathbf{b} = (X^T X)^{-1} X^T \mathbf{Y}$$

  is the unique solution to the normal equations.

Least Squares Estimation

- $\mathbf{b} = (X^T X)^{-1} X^T \mathbf{Y}$ is the Best Linear Unbiased Estimator (BLUE) for β
- BLUE: For any vector of constants $a^T = (a_1, a_2, \ldots, a_{k+1})$,

$$Var(a^T \mathbf{b}) = a^T Var(\mathbf{b}) a$$

is no larger than $Var(a^T b^*)$ for any other linear, unbiased estimator b^* for β

Least Squares Estimation

$$
\begin{aligned}
E(\mathbf{b}) &= E((X^T X)^{-1} X^T \mathbf{Y}) \\
&= (X^T X)^{-1} X^T E(\mathbf{Y}) \\
&= (X^T X)^{-1} X^T X \beta \\
&= \boldsymbol{\beta}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(\mathbf{b}) &= \mathrm{Var}((X^T X)^{-1} X^T \mathbf{Y}) \\
&= (X^T X)^{-1} X^T \mathrm{Var}(\mathbf{Y}) X (X^T X)^{-1} \\
&= (X^T X)^{-1} X^T (\sigma^2 I) X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1}
\end{aligned}
$$

Least Squares Estimation

- The derivation of $Var(\mathbf{b}) = \sigma^2(X^T X)^{-1}$
  - Required uncorrelated errors
  - Required homogeneous error variances
  - Did not require a normal distribution for the random errors (normality is needed for inference procedures)
- An unbiased estimator for σ^2 is

$$s_e^2 = MS_{\text{error}} = \frac{(\mathbf{Y} - X\mathbf{b})^T(\mathbf{Y} - X\mathbf{b})}{n - (k+1)} = \frac{e^T e}{df_{\text{error}}} = \frac{\sum e_i^2}{df_{\text{error}}}$$

- Estimate $\text{Var}(\mathbf{b}) = \sigma^2(X^T X)^{-1}$ as $MS_{\text{error}}(X^T X)^{-1}$

Least Squares Estimation

- $\hat{Y}_i = \mathbf{x}_i^T \mathbf{b} = b_o + b_1 x_{i1} + \cdots + b_k x_{ik}$ is the fitted value or predicted value
- Then

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = X\mathbf{b} = X(X^T X)^{-1} X^T \mathbf{Y} = P_X \mathbf{Y}$$

where $P_X = X(X^T X)^{-1} X^T$ is the orthogonal projection matrix (the perpendicular projection operator) that projects $\mathbf{Y}$ onto the column space of matrix X

Least Squares Estimation

- Given $\hat{\mathbf{Y}} = X\mathbf{b} = P_X\mathbf{Y}$, $e_i = Y_i - \hat{Y}_i$ is the i^{th} residual
- Then $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - P_X\mathbf{Y} = (I - P_X)\mathbf{Y}$
- The matrix $I - P_X$ projects Y onto the space orthogonal to the column space of X (the residual space) as $P_X(I - P_X) = \mathbf{o}$

ANOVA

- Total variability in response variable

$$SS_{\text{Total}} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- Total variability explained by the model

$$SS_{\text{model}} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

- Total variability not explained by the model

$$SS_{\text{error}} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

ANOVA

- Partition the corrected total sum of squares as

$$\begin{aligned} SS_{\text{Total}} &= \sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2 \\ &= SS_{\text{error}} + SS_{\text{model}} \end{aligned}$$

- This partitioning is also expressed as

$$Y^T(I - P_{\mathbf{1}})Y = Y^T(I - P_X)Y + Y^T(P_X - P_{\mathbf{1}})Y$$

where $P_{\mathbf{1}} = P_X$ with $X = [1\ 1\ 1\ \cdots\ 1]^T$

ANOVA Table

source of variation	degrees of freedom	sums of squares
model	k	$SS_{\text{model}} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
error	$n-(k+1)$	$SS_{\text{error}} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
Total	$n-1$	$SS_{\text{Total}} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Estimated Error Variance

$$MS_{\text{error}} = \frac{SS_{\text{error}}}{n - (k+1)}$$

- $E(MS_{\text{error}}) = \sigma^2$ (unbiased estimator)
- $s_e = \sqrt{MS_{\text{error}}}$

Estimated Model Variance

$$MS_{\text{model}} = \frac{SS_{\text{model}}}{k}$$

- $E(MS_{\text{model}}) = \sigma^2 + \dfrac{\beta^T X^T (P_X - P_1) X\beta}{k}$
- If at least one of the $\beta_j \neq 0, j = 1, \ldots, k$,

$$E(MS_{\text{model}}) > \sigma^2$$

F-test for Significance of Model

- $H_o : \beta_1 = \beta_2 = \cdots = \beta_k = 0$
- H_a : at least one $\beta_j \neq 0, j = 1, \ldots, k$
- Test Statistic:
$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}$$
- Reject H_0 if $F > F_{k,n-(k+1),1-\alpha}$
- F-test from ANOVA Table is comparing two models:
  - Model under H_0: $Y_i = \beta_0 + \epsilon_i$
  - Model under H_a: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$
- We almost always reject H_0 in this test

Coefficient of Determination

$$R^2 = \frac{SS_{\text{model}}}{SS_{\text{Total}}}$$

- Fraction of variation in the response variable that can be explained by the multiple linear regression model
- Expressed as percentage: $0\% \leq R^2 \leq 100\%$
- Adding explanatory variables to the model will always increase the value of R^2

Adjusted R^2

$$\text{adj } R^2 = 1 - \frac{MS_{\text{error}}}{SS_{\text{Total}}/(n-1)}$$

- Expressed as percentage: $0\% \leq \text{adj } R^2 \leq 100\%$
- Adjusts for the number of explanatory variables in model through degrees of freedom of $MS_{\text{error}} = n - (k+1)$
- Used primarily for model comparisons

Inference for Population Coefficients

- Test for significance of x_j in model with other explanatory variables
- Two approaches
  - t-test for coefficient
  - Effect test (F-test)
- Results are equivalent

Inference for Population Coefficient

- Least squares estimate for β is $\mathbf{b} = (X^T X)^{-1} X^T \mathbf{Y}$
- Any particular b_j is a linear combination of the elements of the vector $\mathbf{Y}$.
- Y_i are normal random variables, meaning that

$$b_j \text{ is } N(\beta_j, \sigma^2 (X^T X)^{-1}_{[j+1, j+1]})$$

where the variance is the $[j+1, j+1]$ element of the matrix $\sigma^2 (X^T X)^{-1}$

Hypothesis Test for Population Coefficient

- Null and Alternative Hypotheses
  $H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$
- Test Statistic

$$T = \frac{b_j - 0}{s_e \sqrt{(X^T X)^{-1}_{[j+1,j+1]}}} = \frac{b_j - 0}{S_{b_j}}$$

- Reject H_0 if $|T| > t_{n-(k+1),1-\alpha/2}$

Hypothesis Test for Population Coefficients

- Model under H_0

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{j-1} x_{i,j-1} + \beta_{j+1} x_{i,j+1} + \cdots + \beta_k x_{ik} + \epsilon_i$$

- Model under H_a

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{j-1} x_{i,j-1} + \beta_j x_{ij} + \beta_{j+1} x_{i,j+1} + \cdots + \beta_k x_{ik} + \epsilon_i$$

- Significance test for x_j depends on presence or absence of other explanatory variables in model

Confidence Interval for Population Coefficient

- $100(1-\alpha)\%$ CI for β_j is

$$b_j \pm t_{n-(k+1),1-\alpha/2} S_{b_j}$$

Effect Test for Population Coefficient

- Fit two models
  - Model without x_j
  - Model with x_j
- Compare SS_{error} for both models
  - Reduced model without x_j: SSE_{reduced}
  - Full model with x_j: SSE_{full}

Effect Test for Population Coefficient

$$SSE_{\text{reduced}} - SSE_{\text{full}}$$

- Amount of error explained by adding x_j to the model
- The only difference in these two models is the explanatory variable x_j
- Difference has 1 d.f.
- Compare amount of error explained to MSE_{full}

$$F = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/1}{MSE_{\text{full}}}$$

- Large values of F indicate explanatory variable x_j should be included in the model

Effect Test for Population Coefficient

- Null and Alternative Hypotheses
  $H_o : \beta_j = 0 \qquad H_a : \beta_j \neq 0$
- Test Statistic

$$F = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/1}{MSE_{\text{full}}}$$

- Decision - Reject H_o if $F > F_{1,n-(k+1),1-\alpha}$
- Conclusion about x_j is based on other explanatory variables in the model

Partial F-Test

Effect test for significance of a group of m explanatory variables in the model

- Fit two models
  - Reduced Model without the m explanatory variables (only other $k-m$ explanatory variables)
  - Full Model with the m explanatory variables (plus other $k-m$ explanatory variables)
- Compare SS_{error} for both models
  - Reduced model without m explanatory variables: SSE_{reduced}
  - Full model with m explanatory variables: SSE_{full}

Partial F-Test

$$SSE_{\text{reduced}} - SSE_{\text{full}}$$

- Amount of error explained by adding the *m* explanatory variables to the model
- The only difference in these two models is the *m* explanatory variables
- Difference has *m* d.f.
- Compare amount of error explained to MSE_{full}

$$F = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/m}{MSE_{\text{full}}}$$

- Large values of *F* indicate group of *m* explanatory variables should be included in the model

Partial F-Test

- $H_0 : \beta_j = 0$ for the m explanatory variables
- $H_a :$ at least one $\beta_j \neq 0$ for the m explanatory variables
- Test Statistic

$$F = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/m}{MSE_{\text{full}}}$$

- Decision: Reject H_0 if $F > F_{m,n-(k+1),1-\alpha}$
- Conclusion about the significance of the m explanatory variables depends on the presence of the other $k - m$ explanatory variables in the model.

Inference for Conditional Means

Estimate the conditional mean response $\mu_{Y|\mathbf{x}}$ under specific values for vector $\mathbf{x} = (1, x_1, x_2, \ldots, x_k)^T$

- Point estimate is $\hat{\mu}_{Y|\mathbf{x}} = \mathbf{x}^T \hat{\boldsymbol{\beta}}$
- Std error is $S_{\hat{\mu}_{Y|\mathbf{x}}} = \sqrt{MS_{\text{error}}\, \mathbf{x}^T (X^T X)^{-1} \mathbf{x}}$
- A $(1 - \alpha) \times 100\%$ confidence interval for $\mu_{Y|\mathbf{x}}$ is

$$\hat{\mu}_{Y|\mathbf{x}} \pm t_{n-(k+1), 1-\alpha/2}\, S_{\hat{\mu}_{Y|\mathbf{x}}}$$

- Simultaneous confidence region for an entire line segment (the Scheffe's method) is

$$\hat{Y} \pm \sqrt{(k+1) F_{k+1, n-k-1, 1-\alpha}}\, S_{\hat{\mu}_{Y|\mathbf{x}}}$$

Prediction Intervals

Predict value of $Y_i = \mathbf{x}^T\beta + \epsilon_i$ that will be observed under specific values for vector $\mathbf{x} = (1, x_1, x_2, \ldots, x_k)^T$

- The predictor is $\hat{Y}_i = \mathbf{x}^T\hat{\beta}$
- The standard error for the predictor is $S_{\hat{Y}} = \sqrt{MS_{\text{error}} + S^2_{\hat{\mu}_{Y|\mathbf{x}}}}$
- A $(1-\alpha) \times 100\%$ prediction interval is

$$\hat{Y}_i \pm t_{n-(k+1),1-\alpha/2} S_{\hat{Y}}$$

Unit 3

Multiple Linear Regression (MLR)

Examples

Grandfather Clock Example

- There were 32 antique ($>$100 years old) grandfather clocks sold at auction
- Response variable: price at auction
- Two explanatory variables:
  - Age (in years)
  - Number of bidders

Grandfather Clock: Data

Price Y	Age (years) X_1	NumBid X_2
1235	127	13
1080	115	12
845	127	7
.	.	.
.	.	.
.	.	.
1262	168	7

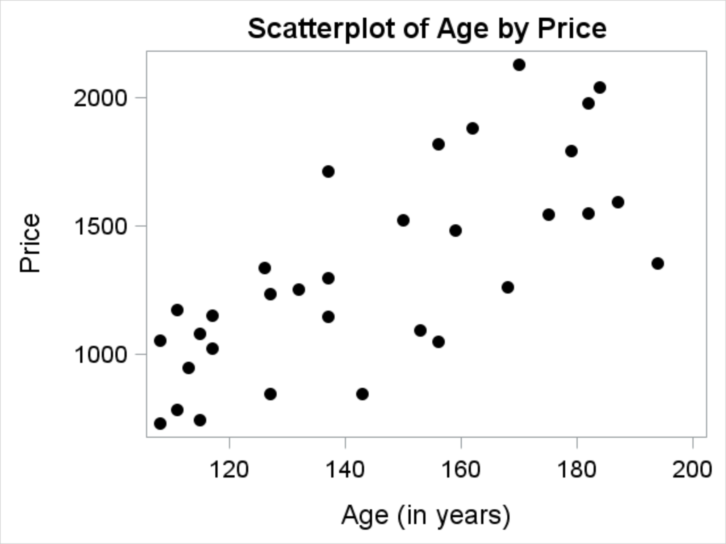
Grandfather Clock: Different Regression Analysis

$$\hat{Y}_i = \beta_0 + \beta_1 \, Age$$

$$\hat{Y}_i = \beta_0 + \beta_2 \, NumBid$$

$$\hat{Y}_i = \beta_0 + \beta_1 \, Age + \beta_2 \, NumBid$$

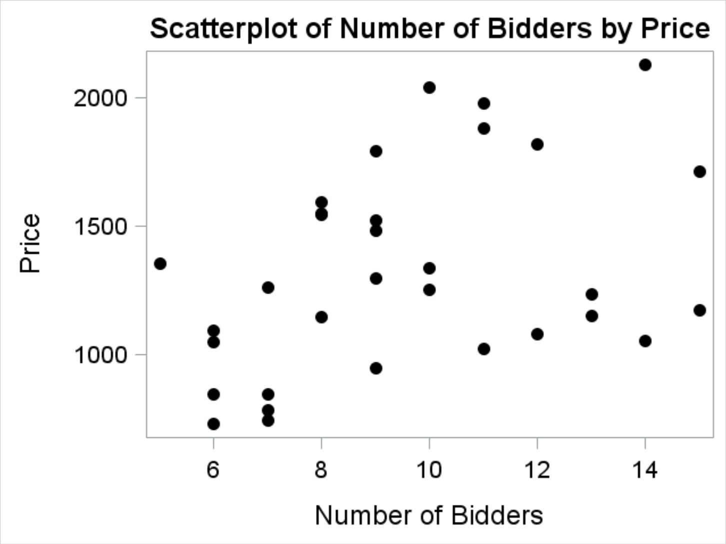
MLR: Examples



Grandfather Clock: SLR of Age on Price

$$\hat{Y}_i = -192.05 + 10.48\, Age$$

- There is a significant linear relationship between age and price at auction (F-test p-value <0.0001)
- Each additional year of age is associated with a mean increase in price of 10.48 dollars
- $R^2 = 53.24\%$ of the variation in price can be explained by the linear regression model with age


Scatterplot of Number of Bidders by Price
Price
2000
1500
1000
6
8
10
12
14
Number of Bidders

Grandfather Clock: SLR of Number of Bidders on Price

$$\hat{Y}_i = 804.91 + 54.76\ NumBid$$

- There is a significant linear relationship between number of bidders and price at auction (F-test p-value=0.0252)
- Each additional additional bidder is associated with a mean increase in price of 54.76 dollars
- $R^2 = 15.62\%$ of the variation in price can be explained by the linear regression model with number of bidders

Grandfather Clock: MLR on Price

With both explanatory variables in the MLR, the dimension of the design matrix X is 32×3.

$$X = \begin{bmatrix} 1 & 127 & 13 \\ 1 & 115 & 12 \\ 1 & 127 & 7 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 168 & 7 \end{bmatrix}$$

Grandfather Clock: MLR on Price

With both explanatory variables in the MLR, the dimension of the estimated coefficient vector **b** is 1×3.

$$\mathbf{b} = (X^T X)^{-1} X^T \mathbf{Y} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -1338.95 \\ 12.74 \\ 85.95 \end{bmatrix}$$

Estimated Regression Model:
$\hat{Y}_i = -1338.95 + 12.74\, Age + 85.95\, NumBid$

Grandfather Clock: Different Regression Analysis

$$\hat{Y}_i = -192.05 + 10.48\, Age$$

$$\hat{Y}_i = 804.91 + 54.76\, NumBid$$

$$\hat{Y}_i = -1338.95 + 12.74\, Age + 85.95\, NumBid$$

Grandfather Clock: MLR on Price

$$
\begin{aligned}
MS_{error}(X^TX)^{-1} &= 17818\begin{bmatrix} 1.695 & -0.00773 & -0.057 \\ -0.00773 & 0.0000459 & 0.0001 \\ -0.057 & 0.0001 & 0.00428 \end{bmatrix} \\
&= \begin{bmatrix} 30209 & -137.74 & -1016.58 \\ -137.74 & 0.8185 & 2.004 \\ -1016.58 & 2.004 & 76.186 \end{bmatrix}
\end{aligned}
$$

Then

$$
\begin{aligned}
S_{b_0} &= \sqrt{30209} = 173.81 \\
S_{b_1} &= \sqrt{0.8185} = 0.9047 \\
S_{b_2} &= \sqrt{76.186} = 8.7285
\end{aligned}
$$

Grandfather Clock: Confidence Interval for β_1

- β_1 represents the change in auction price when age is increased 1 year while the number of bidders is held constant.
- A $(1-\alpha) \times 100\%$ confidence interval for β_1:

$$b_1 \pm t_{df_{error}, 1-\alpha/2}\, S_{b_1}$$

- A 95% confidence interval is

$$12.74 \pm (2.045)(0.9047) \quad \Rightarrow \quad (10.89, 14.59)$$

Grandfather Clock: Hypothesis Test for β_1

- $H_o : \beta_1 = 0$ or $E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_2 x_2$

versus

- $H_a : \beta_1 \neq 0$ or $E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- Test statistic:

$$t = \frac{b_1 - 0}{S_{b_1}} = \frac{12.74}{0.9047} = 14.08$$

on 29 df with p-value < 0.0001.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4283063	2141531	120.19	<.0001
Error	29	516727	17818		
Corrected Total	31	4799790			

Root MSE	133.48467	R-Square	0.8923
Dependent Mean	1326.87500	Adj R-Sq	0.8849
Coeff Var	10.06008		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-1338.95134	173.80947	-7.70	<.0001	-1694.43162	-983.47106
age	1	12.74057	0.90474	14.08	<.0001	10.89017	14.59098
numbid	1	85.95298	8.72852	9.85	<.0001	68.10115	103.80482

Grandfather Clock: MLR on Price

$$\hat{Y}_i = -1338.95 + 12.74\, Age + 85.95\, NumBid$$

- Model is statistically significant in explaining Price with $F = 120.9$ and p-value < 0.0001.
- $R^2 = 89.23\%$ of the variation in price can be explained by the multiple linear regression model with both age and number of bidders
- Given number of bidders in the model, age is statistically significant with $t = 14.08$ and p-value < 0.0001
- Given age in the model, number of bidders is statistically significant with $t = 9.85$ and p-value < 0.0001

Grandfather Clock: MLR on Price

$$\hat{Y}_i = -1338.95 + 12.74\, Age + 85.95\, NumBid$$

- This analysis indicates that changes in either Age (X_1) or Number of Bidders (X_2) affect the auction price.
  - Holding the number of bidders constant, a 1 year increase in age increases price by 12.74 dollars.
  - Holding age constant, a 1 additional bidder increase increases auction price by 85.95 dollars.
  - What if you change both age and number of bidders?
  - How should the intercept be interpreted?
- The significance of each coefficient does not necessarily imply that the model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$ is correct

Grandfather Clock: MLR on Price

Estimate the mean price of a clock when

X_1 = Age = 150 years
X_2 = NumBid = 10

In this case

$$x^T = (1 \;\; 150 \;\; 10)$$

The least squares estimate of the mean yield under these conditions is

$$\hat{Y} = x^T b = (1 \;\; 150 \;\; 10) \begin{bmatrix} -1338.95 \\ 12.74 \\ 85.95 \end{bmatrix} = 1431.55$$

Grandfather Clock: MLR on Price

Compute the standard error of the estimated mean

$$
\begin{aligned}
S^2_{\hat{Y}} &= MS_{error} x^T (X^T X)^{-1} x \\
&= x^T \left[MS_{error} (X^T X)^{-1} \right] x \\
&= (1\ 150\ 10) \begin{bmatrix} 30209 & -137.74 & -1016.58 \\ -137.74 & 0.8185 & 2.004 \\ -1016.58 & 2.004 & 76.186 \end{bmatrix} \begin{bmatrix} 1 \\ 150 \\ 10 \end{bmatrix} \\
&= 604.04
\end{aligned}
$$

The standard error is $S_{\hat{Y}} = \sqrt{604.04} = 24.58$

Grandfather Clock: MLR on Price

- A $(1-\alpha)\times 100\%$ confidence interval for the mean price under the conditions specified by $x = (1\ 150\ 10)$ is

$$\hat{Y} \pm t_{df_{error},1-\alpha/2}\, S_{\hat{Y}}$$

- A 95% confidence interval is

$$1431.55 \pm (2.045)(24.58) \quad \Rightarrow \quad (1381.28, 1481.82)$$

Grandfather Clock: MLR on Price

Predict price of a clock to be sold at a future auction when

X_1 = Age = 150 years
X_2 = NumBid = 10

In this case

$$x^T = (1 \;\; 150 \;\; 10)$$

The predicted value of the random error is zero and the predicted price under the conditions specified by x is

$$\hat{Y} = x^T b + 0 = (1 \;\; 150 \;\; 10) \begin{bmatrix} -1338.95 \\ 12.74 \\ 85.95 \end{bmatrix} = 1431.55$$

Grandfather Clock: MLR on Price

Compute the standard error of the predicted price

$$\begin{aligned} S^2_{pred} &= MS_{error} + MS_{error} x^T (X^T X)^{-1} x \\ &= MS_{error} + S^2_{\hat{Y}} \\ &= 17818 + 604.04 \\ &= 18422.04 \end{aligned}$$

The standard error is

$$S_{pred} = \sqrt{18422.04} = 135.73$$

Grandfather Clock: MLR on Price

- $(1 - \alpha) \times 100\%$ prediction interval for the price under the conditions specified by $x = (1\ 150\ 10)$ is

$$\hat{Y} \pm t_{df_{error},1-\alpha/2}\, S_{pred}$$

- A 95% prediction interval is

$$1431.55 \pm (2.045)(135.73) \quad \Rightarrow \quad (1153.98, 1709.12)$$

Grandfather Clock: MLR on Price

- $(1-\alpha)\times 100\%$ simultaneous prediction region for the auction price

$$\hat{Y} \pm \sqrt{(k+1)F_{(k+1,df_{error}),1-\alpha}}\, S_{pred}$$

- Simultaneous 95% prediction intervals are

$$\hat{Y} \pm \sqrt{3F_{(3,29),0.95}} S_{pred}$$

$\Rightarrow$

$$\hat{Y} \pm \sqrt{(3)(2.934)} S_{pred}$$

$\Rightarrow$

$$\hat{Y} \pm (2.9668) S_{pred}$$

Grandfather Clock: Effect Test for β_2 (NumBid)

Source	d.f.	SS	MS	F	p-val
Model with Age	1	2555224	2555224	34.15	< 0.0001
Error	30	2244565	74819		
corrected total	31	4799790			

Source	d.f.	SS	MS	F	p-val
Model with Age and NumBid	2	4283063	2141531	120.19	< 0.0001
Error	29	516727	17818		
corrected total	31	4799790			

Grandfather Clock: Effect Test for β_2 (NumBid)

- Adding Number of Bidders to the SLR model with Age reduces the SS_{Error} for the model
- For SLR with Age, SS_{Error} = 2244565
- For MLR with Age and NumBid, SS_{Error} = 516727
- Difference = 2244565 - 516727 = 1727838

Grandfather Clock: Effect Test for β_2 (NumBid)

$$
\begin{aligned}
F &= \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/m}{MSE_{\text{full}}} \\
&= \frac{(SSE_{\text{SLR Age}} - SSE_{\text{MLR}})/m}{MSE_{\text{MLR}}} \\
&= \frac{1727838/1}{17818} \\
&= 96.97 \\
&= 9.85^2
\end{aligned}
$$

Grandfather Clock: MLR on Price

$$\hat{Y}_i = -1338.95 + 12.74\, Age + 85.95\, NumBid$$

- This model is additive
  - The effect of age on the price of a clock is the same for each number of bidders
  - The effect of number of bidders on the price of a clock is the same for every value of age

Grandfather Clock: MLR with Interaction

- Allows for the effect of one explanatory variable on the response variable to be different depending on the value of another explanatory variable.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$$

- Effect on Response Variable
  - The effect of increasing x_{i1} by 1 is $\beta_1 + \beta_3 x_{i2}$.
  - The effect of increasing x_{i2} by 1 is $\beta_2 + \beta_3 x_{i1}$.

MLR: Examples

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	4578427	1526142	193.04	<.0001
Error	28	221362	7905.79047		
Corrected Total	31	4799790			

Root MSE	88.91451	**R-Square**	0.9539
Dependent Mean	1326.87500	**Adj R-Sq**	0.9489
Coeff Var	6.70105		

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	320.45799	295.14128	1.09	0.2868	-284.11152	925.02751
age	1	0.87814	2.03216	0.43	0.6690	-3.28454	5.04083
numbid	1	-93.26482	29.89162	-3.12	0.0042	-154.49502	-32.03462
agexnumbid	1	1.29785	0.21233	6.11	<.0001	0.86290	1.73279

Test for Significant Interaction

- T-test: $t = 6.11$, p-value < 0.0001
- Effect test:
  - For MLR with Age and NumBid, SS_{Error} = 516727
  - For MLR with Age and NumBid and interaction, SS_{Error} = 221362

Test for Significant Interaction

- The partial F-test:

$$\begin{aligned} F &= \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/m}{MSE_{\text{full}}} \\ &= \frac{(516727 - 221362)/1}{7905.79} \\ &= 37.36 \\ &= 6.11^2 \end{aligned}$$

- Interaction Term is statistically significant in model

Tests for Component Explanatory Variables

- Do not perform significance tests for component explanatory variables when corresponding interaction terms exist in the model
- These tests no longer have any meaning
  - Test for significance of variable given the other variables in the model
  - The component variable is already in the model through its presence in the interaction term
  - Cannot separate significance of component variable from its interaction term

Alternative Parameterization of Interaction Term

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) + \epsilon_i \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} - \beta_3 x_{i1} \bar{x}_2 - \beta_3 x_{i2} \bar{x}_1 + \beta_3 \bar{x}_1 \bar{x}_2 + \epsilon_i \\ &= \beta_0 + \beta_3 \bar{x}_1 \bar{x}_2 + (\beta_1 - \beta_3 \bar{x}_2) x_{i1} + (\beta_2 - \beta_3 \bar{x}_1) x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i \end{aligned}$$

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	4578427	1526142	193.04	<.0001
Error	28	221362	7905.79047		
Corrected Total	31	4799790			

Root MSE	88.91451	**R-Square**	0.9539
Dependent Mean	1326.87500	**Adj R-Sq**	0.9489
Coeff Var	6.70105		

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	-1472.43236	117.81657	-12.50	<.0001	-1713.76866	-1231.09606
age	1	13.24824	0.60835	21.78	<.0001	12.00209	14.49438
numbid	1	94.84170	5.99320	15.82	<.0001	82.56519	107.11822
cagexcnumbid	1	1.29785	0.21233	6.11	<.0001	0.86290	1.73279

Alternative Parameterization of Interaction Term

- Estimated coefficient for interaction term does not change
- Estimated coefficients for intercept and component explanatory variables change
  - Different std. errors, t-test statistics and p-values
- Correlation between component explanatory variables and interaction term is reduced

Uncorrelated Predictors

Example: Yield of a chemical process (Myers)

Y = Yield (%)
X_1 = Temperature (^{o}F)
X_2 = Time (hours)

Data:

Y	X_1	X_2
77	160	1
79	160	2
82	165	1
83	165	2
85	170	1
88	170	2
90	175	1
93	175	2

Chemical Process Study

Full Factorial Design

$$r_{x_1,x_2} = \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2}} = 0$$

Estimated Models

- Model 1: $\hat{\mathbf{Y}}_i = -64.45 + 0.890x_{i_1}$

$$R^2 = 0.9435$$

- Model 2: $\hat{\mathbf{Y}}_i = 81.25 + 2.25x_{i_2}$

$$R^2 = 0.0482$$

- Model 12: $\hat{\mathbf{Y}}_i = -67.825 + 0.890x_{i_1} + 2.250x_{i_2}$

$$R^2 = 0.9918$$

Chemical Process Study

Source of variation	d.f.	SS	MS	F	p-val
reg on x_1	1	198.025	198.025	574.0	.0001
reg on x_2 after x_1	1	10.125	10.125	29.3	.0029
error	5	1.725	0.345		
corrected total	7	209.875			

Source of variation	d.f.	SS	MS	F	p-val
reg on x_2	1	10.125	10.125	29.3	.0029
reg on x_1 after x_2	1	198.025	198.025	574.0	.0001
error	5	1.725	0.345		
corrected total	7	209.875			

Complete Confounding

Example: Correlation between X_1 and X_2 is one

Y	X_1	X_2
1.95	1	5
6.25	2	10
9.85	3	15

Estimated Models

- Model 1: $\hat{\mathbf{Y}}_i = -1.8833 + 3.95x_{i1}$

$$R^2 = 0.9974$$

- Model 2: $\hat{\mathbf{Y}}_i = -1.8833 + 0.79x_{i2}$

$$R^2 = 0.9974$$

- Model 12: Many choices for b_1 and b_2 in

$$\begin{aligned}\hat{\mathbf{Y}}_i &= b_0 + b_1x_{i1} + b_2x_{i2} = b_0 + b_1x_{i1} + b_2(5x_{i1}) \\ &= b_0 + (b_1 + 5b_2)x_{i1}\end{aligned}$$

$$R^2 = 0.9974$$

Complete Confounding Example

Source of variation	d.f.	SS	MS	F	p-val
reg on x_1	1	31.205	31.205	382.1	.0325
reg on x_2 after x_1	0	0.000	0.000	NA	NA
error	1	0.08167	0.08167		
corrected total	2	31.28667			

Source of variation	d.f.	SS	MS	F	p-val
reg on x_2	1	31.205	31.205	382.1	.0325
reg on x_1 after x_2	0	0.000	0.000	NA	NA
error	1	0.08167	0.08167		
corrected total	2	31.28667			

Partial Confounding

Example: Correlation between X_1 and X_2 is 0.95237

Y	X_1	X_2
1.8	1.0	5
1.7	1.1	6
5.4	1.8	11
6.1	2.0	10
7.0	2.1	9
9.6	3.0	15

Estimated Models

- Model 1: $\hat{\mathbf{Y}}_i = -2.328 + 4.142x_{i1}$
$$R^2 = 0.978$$
- Model 2: $\hat{\mathbf{Y}}_i = -2.114 + 0.791x_{i2}$
$$R^2 = 0.865$$
- Model 12: $\hat{\mathbf{Y}}_i = -2.247 + 4.655x_{i1} - 0.109x_{i2}$
$$R^2 = 0.980$$

Partial Confounding Example

Source of variation	d.f.	SS	MS	F	p-val
reg on x_1	1	46.215	46.215	146.6	0.0012
reg on x_2 after x_1	1	0.073	0.073	0.23	0.6639
error	3	0.946	0.315		
corrected total	5	47.233			

Source of variation	d.f.	SS	MS	F	p-val
reg on x_2	1	40.859	40.859	129.6	0.0015
reg on x_1 after x_2	1	5.428	5.428	17.21	0.0254
error	3	0.946	0.315		
corrected total	5	47.233			

Interpreting Regression Coefficients

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

- β_j is the jth regression coefficient or the jth *partial* regression coefficient
- β_j is the change in the mean of Y for a unit change in X_j **with all other variables held constant**
- Sometimes this is not possible and the values of other explanatory variables change when X_j changes: (e.g., polynomial terms (X_j, X_j^2) or interaction terms $(X_i, X_j, X_i X_j)$ or other highly correlated predictors)

Interpreting Regression Coefficients

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

- An alternative interpretation: β_j is the linear effect of X_j on Y after adjusting for the linear effect of the other predictors on Y and the linear effects of the other predictors on X_j
- Let P_{-x_j} represent the projection matrix without variable X_j (delete column $j+1$ from the model matrix X). Then, $\hat{\beta}_j$ is found from the regression of $(I - P_{-x_j})Y$ on $(I - P_{-x_j})X_j$

Interpreting Regression Coefficients

- Example: Brain size data (an observational study)
  Question: Do species with longer gestation times have bigger brains?
- Plots, biology $\Rightarrow$ linear in log variables
- Model 1: $log(brain)_i = \beta_0 + \beta_1 log(gest_i) + \epsilon_i$

  $\hat{\beta}_1 = 2.23 \Rightarrow$ Species differing by 1 unit log gestation time (e.g. log(gest) = 2 and log(gest)=1) differ in log(brain size) by 2.23 units, on average.
- Biology $\Rightarrow$ body size associated with both

Interpreting Regression Coefficients

- Model 2:

  $log(brain_i) = \beta_0 + \beta_1 log(gest_i) + \beta_2 log(body_i) + \epsilon_i$

  $\hat{\beta}_1 = 0.668$

  Two species with the same body size but differing by 1 unit log gestation time differ in log brain size by 0.668 units, on average.

- So, when is β_j in multiple regression equal to β_j from simple linear regression?

  Answer: When X_j is uncorrelated with the rest of the explanatory variables.

Interpreting Regression Coefficients

- Consider the regression of one set of residuals $(I - P_{-x_j})Y$ on another set of residuals $(I - P_{-x_j})X_j$
  - Regress log(brain) on log(body):
    residual $= e_i = (I - P_{-x_j})Y$
  - Regress log(gest) on log(body):
    residual$= g_i = (I - P_{-x_j})X_j$
  - β_2 is regression coefficient for regression of e_i on g_i:

$$e_i = \beta_2 g_i + \eta_i$$

Questions?

Contact me:

Email: dmommen@iastate.edu
Student Office Hours: Thursdays @ 10-11 am