Regression and Causal Inference

September 6, 2012

Prediction

- We have an input vector $X^T = (X_1, X_2, \ldots, X_p)$ with dimensions of $n \times p$ and an output vector Y with dimensions $n \times 1$.

Prediction

- We have an input vector $X^T = (X_1, X_2, \ldots, X_p)$ with dimensions of $n \times p$ and an output vector Y with dimensions $n \times 1$.
- The linear regression model has the form:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

Prediction

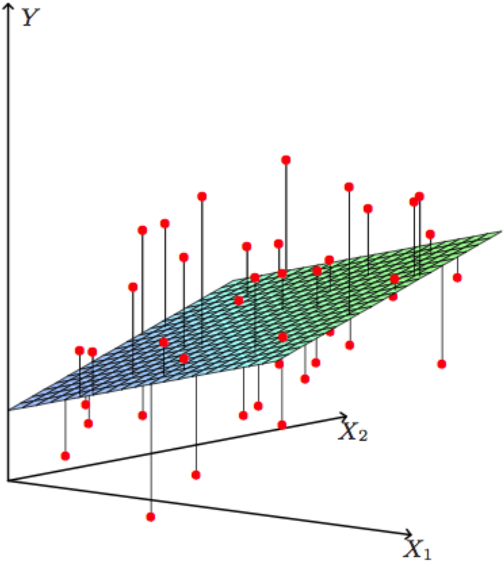
- We have an input vector $X^T = (X_1, X_2, \dots, X_p)$ with dimensions of $n \times p$ and an output vector Y with dimensions $n \times 1$.
- The linear regression model has the form:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

- We can pick the coefficients $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ in a variety of ways but OLS is by far the most common, which minimizes the **residual sum of squares** (RSS):

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 \end{aligned}$$

OLS in a Picture



Deriving the Algorithm

- Denote $\mathbf{X}$ the $N \times (p+1)$ matrix with each row an input vector (with a 1 in the first position) and $\mathbf{y}$ is the output vector.

Deriving the Algorithm

- Denote $\mathbf{X}$ the $N \times (p+1)$ matrix with each row an input vector (with a 1 in the first position) and $\mathbf{y}$ is the output vector.
- Write the RSS as:

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{x}\beta)$$

Deriving the Algorithm

- Denote $\mathbf{X}$ the $N \times (p+1)$ matrix with each row an input vector (with a 1 in the first position) and $\mathbf{y}$ is the output vector.
- Write the RSS as:

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{x}\beta)$$

- Differentiate with respect to β:

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \tag{1}$$

Deriving the Algorithm

- Denote $\mathbf{X}$ the $N \times (p+1)$ matrix with each row an input vector (with a 1 in the first position) and $\mathbf{y}$ is the output vector.
- Write the RSS as:

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{x}\beta)$$

- Differentiate with respect to β:

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \qquad (1)$$

- Assume that $\mathbf{X}$ is full rank (no perfect collinearity among any of the independent variables) and set first derivative to 0:

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

Deriving the Algorithm

- Denote $\mathbf{X}$ the $N \times (p+1)$ matrix with each row an input vector (with a 1 in the first position) and $\mathbf{y}$ is the output vector.
- Write the RSS as:

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{x}\beta)$$

- Differentiate with respect to β:

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \qquad (1)$$

- Assume that $\mathbf{X}$ is full rank (no perfect collinearity among any of the independent variables) and set first derivative to 0:

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

- Solve for β:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Making a Prediction

- The *hat matrix*, or *projection matrix*

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \text{ with } \tilde{\mathbf{H}} = \mathbf{I} - \mathbf{H}$$

Making a Prediction

- The *hat matrix*, or *projection matrix*

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \text{ with } \tilde{\mathbf{H}} = \mathbf{I} - \mathbf{H}$$

- We use the hat matrix to find the fitted values:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

Making a Prediction

- The *hat matrix*, or *projection matrix*

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \text{ with } \tilde{\mathbf{H}} = \mathbf{I} - \mathbf{H}$$

- We use the hat matrix to find the fitted values:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

- We can now write

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Making a Prediction

- The *hat matrix*, or *projection matrix*

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \text{ with } \tilde{\mathbf{H}} = \mathbf{I} - \mathbf{H}$$

- We use the hat matrix to find the fitted values:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

- We can now write

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

- If $\mathbf{HY}$ yields part of $\mathbf{Y}$ that projects into $\mathbf{X}$, this means that $\tilde{\mathbf{H}}\mathbf{Y}$ is the part of $\mathbf{Y}$ that does not project into $\mathbf{X}$, which is the *residual* part of $\mathbf{Y}$. Therefore, $\tilde{\mathbf{H}}\mathbf{Y}$ makes the residuals.

From Algorithm to Model

1. *Linear in Parameters*: Y is related to the independent variables and the error term as $Y = X\beta + \epsilon$

From Algorithm to Model

1. *Linear in Parameters*: Y is related to the independent variables and the error term as $Y = X\beta + \epsilon$
2. The X's are fixed and take on ≥ 2 values

From Algorithm to Model

1. *Linear in Parameters*: Y is related to the independent variables and the error term as $Y = X\beta + \epsilon$
2. The X's are fixed and take on ≥ 2 values
3. *Full Rank* (in multiple regression): There is no perfect collinearity among any of the independent variables

From Algorithm to Model

1. *Linear in Parameters*: Y is related to the independent variables and the error term as $Y = X\beta + \epsilon$
2. The X's are fixed and take on ≥ 2 values
3. *Full Rank* (in multiple regression): There is no perfect collinearity among any of the independent variables
4. *Zero Conditional Mean*: $E(\epsilon|X) = 0$

From Algorithm to Model

1. *Linear in Parameters*: Y is related to the independent variables and the error term as $Y = X\beta + \epsilon$
2. The X's are fixed and take on ≥ 2 values
3. *Full Rank* (in multiple regression): There is no perfect collinearity among any of the independent variables
4. *Zero Conditional Mean*: $E(\epsilon|X) = 0$
5. *Homoskedasticity*: $Var(\epsilon|X) = \sigma^2$

From Algorithm to Model

1. *Linear in Parameters*: Y is related to the independent variables and the error term as $Y = X\beta + \epsilon$
2. The X's are fixed and take on ≥ 2 values
3. *Full Rank* (in multiple regression): There is no perfect collinearity among any of the independent variables
4. *Zero Conditional Mean*: $E(\epsilon|X) = 0$
5. *Homoskedasticity*: $Var(\epsilon|X) = \sigma^2$
6. *Random Sampling*: Y_i is an *iid* random sample, although this can be relaxed to
   $cov(y_i, y_j) = 0 = cov(\epsilon_i, \epsilon_j) \qquad i \neq j$

From Algorithm to Model

1. *Linear in Parameters*: Y is related to the independent variables and the error term as $Y = X\beta + \epsilon$
2. The X's are fixed and take on ≥ 2 values
3. *Full Rank* (in multiple regression): There is no perfect collinearity among any of the independent variables
4. *Zero Conditional Mean*: $E(\epsilon|X) = 0$
5. *Homoskedasticity*: $Var(\epsilon|X) = \sigma^2$
6. *Random Sampling*: Y_i is an *iid* random sample, although this can be relaxed to $cov(y_i, y_j) = 0 = cov(\epsilon_i, \epsilon_j) \qquad i \neq j$
7. *Normal Errors* (optional): $Y \sim \mathbb{N}(X\beta, \sigma^2)$

Gauss-Markov

- Under Assumptions 1-7 above, $\hat{\beta}$ is the best linear unbiased estimator (BLUE) of β.
- **B** est means *smallest variance* amongst linear unbiased estimates,
- **L** inear means $\hat{\beta}$ is estimable from a linear function of the data,
- **U** nbiased means $E(\hat{\beta}) = \beta$,
- **E** stimator means X is full rank.

Unbiasedness

Recall:

$$\begin{aligned}\hat{\beta} &= (X^TX)^{-1}X^TY \\ &= (X^TX)^{-1}X^T(X\beta+\epsilon) \\ &= (X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\epsilon \\ &= \beta + (X^TX)^{-1}X^T\epsilon\end{aligned}$$

We know that $\hat{\beta}$ is unbiased if $E(\hat{\beta}) = \beta$

$$\begin{aligned}E(\hat{\beta}) &= E(\beta + (X^TX)^{-1}X^T\epsilon|X) \\ &= E(\beta|X) + E((X^TX)^{-1}X^T\epsilon|X) \\ &= \beta + (X^TX)^{-1}E(\epsilon|X) \\ &\qquad \text{where } E(\epsilon|X) = E(\epsilon) = 0 \\ E(\hat{\beta}) &= \beta\end{aligned}$$

Deriving σ^2

- Recall:

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y \\ &= (X^T X)^{-1} X^T (X\beta + \epsilon) \\ \Rightarrow \quad \hat{\beta} - \beta &= (X^T X)^{-1} X^T \epsilon \end{aligned}$$

Deriving σ^2

- Recall:

$$
\begin{aligned}
\hat{\beta} &= (X^TX)^{-1}X^TY \\
&= (X^TX)^{-1}X^T(X\beta+\epsilon) \\
\Rightarrow \quad \hat{\beta}-\beta &= (X^TX)^{-1}X^T\epsilon
\end{aligned}
$$

- Plugging this into the covariance equation:

$$
\begin{aligned}
cov(\hat{\beta}|X) &= E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X] \\
&= E\left[\left((X^TX)^{-1}X^T\epsilon\right)\left((X^TX)^{-1}X^T\epsilon\right)'|X\right] \\
&= E[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}|X] \\
&= (X^TX)^{-1}X^TE(\epsilon\epsilon^T|X)X(X^TX)^{-1} \\
&\qquad \text{where } E(\epsilon\epsilon^T|X) = \sigma^2 I_{p\times p} \\
&= (X^TX)^{-1}X^T\sigma^2 I_{p\times p}X(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}
$$

Deriving σ^2

We estimate σ^2 dividing the residuals squared by the degrees of freedom because the e_i are generally smaller than the ϵ_i due to the fact that $\hat{\beta}$ was chosen to make the sum of square residuals as small as possible.

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} e_i^2$$

What Makes OLS the Best?

- We want an estimator $\tilde{\beta} = m + MY$, with $E(\tilde{\beta}|X) = \beta$

$$\begin{aligned} E(\tilde{\beta}|X) &= E(m + MY|X) \\ &= E(m + M(X\beta + \epsilon)|X) \\ &= m + MX\beta \end{aligned}$$

$$\Rightarrow m = 0 \text{ and } MX = I_{p \times p} \tag{2}$$

What Makes OLS the Best?

- We want an estimator $\tilde{\beta} = m + MY$, with $E(\tilde{\beta}|X) = \beta$

$$\begin{aligned} E(\tilde{\beta}|X) &= E(m + MY|X) \\ &= E(m + M(X\beta + \epsilon)|X) \\ &= m + MX\beta \end{aligned}$$

$$\Rightarrow m = 0 \text{ and } MX = I_{p \times p} \qquad (2)$$

- Therefore, this implies we want $\tilde{\beta} = MY$, so WLOG we can say $M = (X^T X)^{-1} X^T + c$

What Makes OLS the Best?

- We want an estimator $\tilde{\beta} = m + MY$, with $E(\tilde{\beta}|X) = \beta$

$$\begin{aligned} E(\tilde{\beta}|X) &= E(m + MY|X) \\ &= E(m + M(X\beta + \epsilon)|X) \\ &= m + MX\beta \end{aligned}$$

$$\Rightarrow m = 0 \text{ and } MX = I_{p\times p} \tag{2}$$

- Therefore, this implies we want $\tilde{\beta} = MY$, so WLOG we can say $M = (X^T X)^{-1} X^T + c$
- Thus,

$$\begin{aligned} MX &= ((X^T X)^{-1} X^T + c)X \\ &= (X^T X)^{-1} X^T X + cX \\ &= I_{p\times p} + CX = I_{p\times p} \text{ by (2)} \end{aligned}$$

$$\Rightarrow CX = 0 \tag{3}$$

What Makes OLS the Best?

- Also note that

$$\tilde{\beta} = MY = M(X\beta + \epsilon) = \beta + M\epsilon \text{ by } MX = I_{p \times p}$$

$$\Rightarrow \tilde{\beta} - \beta = M\epsilon \tag{4}$$

- Now, recall "best" means having the smallest variance, therefore we want to minimize $cov(\tilde{\beta}|X)$

$$\begin{aligned} cov(\tilde{\beta}|X) &= E((\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^T|X) \\ &= E((M\epsilon)(M\epsilon)^T|X) \\ &= E(M\epsilon\epsilon^T M^T|X) \\ &= ME(\epsilon\epsilon^T|X)M^T \\ &= \sigma^2 MM^T \end{aligned}$$

What Makes OLS the Best?

- Finally,

$$\begin{aligned} MM^T &= \left((X^TX)^{-1}X^T + c\right)\left((X^TX)^{-1}X^T + c\right)^T \\ &= (X^TX)^{-1} + CC^T \end{aligned}$$

$$\text{since from (3)}, CX = 0 \text{ and } C^TX^T = 0 \tag{5}$$

What Makes OLS the Best?

- Finally,

$$\begin{aligned} MM^T &= \left((X^TX)^{-1}X^T + c\right)\left((X^TX)^{-1}X^T + c\right)^T \\ &= (X^TX)^{-1} + CC^T \end{aligned}$$

$$\text{since from (3)}, CX = 0 \text{ and } C^TX^T = 0 \tag{5}$$

- Then,

$$cov(\tilde{\beta}|X) = \sigma^2(X^TX)^{-1} + \sigma^2 CC^T$$

What Makes OLS the Best?

- Finally,

$$\begin{aligned} MM^T &= \left((X^TX)^{-1}X^T + c\right)\left((X^TX)^{-1}X^T + c\right)^T \\ &= (X^TX)^{-1} + CC^T \end{aligned}$$

$$\text{since from (3)}, CX = 0 \text{ and } C^TX^T = 0 \qquad (5)$$

- Then,

$$cov(\tilde{\beta}|X) = \sigma^2(X^TX)^{-1} + \sigma^2 CC^T$$

- When is this minimized?

Regression Anatomy

- In the simple bivariate case:

$$\beta_1 = \frac{\mathrm{Cov}(Y_i, X_i)}{\mathrm{Var}(X_i)}$$

Regression Anatomy

- In the simple bivariate case:

$$\beta_1 = \frac{\mathrm{Cov}(Y_i, X_i)}{\mathrm{Var}(X_i)}$$

- In the multivariate case, β_j is:

$$\beta_j = \frac{\mathrm{Cov}(Y_i, \tilde{X}_{ij})}{\mathrm{Var}(\tilde{X}_{ij})}$$

  where $\tilde{X}_{ij}$ is the residual from the regression of X_{ij} on all other covariates.

Regression Anatomy

- In the simple bivariate case:

$$\beta_1 = \frac{\mathrm{Cov}(Y_i, X_i)}{\mathrm{Var}(X_i)}$$

- In the multivariate case, β_j is:

$$\beta_j = \frac{\mathrm{Cov}(Y_i, \tilde{X}_{ij})}{\mathrm{Var}(\tilde{X}_{ij})}$$

  where $\tilde{X}_{ij}$ is the residual from the regression of X_{ij} on all other covariates.

- The multiple regression coefficient $\hat{\beta}_j$ represents the additional contribution of x_j on y, after x_j has been adjusted for $x_o, x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$

Regression Anatomy

- In the simple bivariate case:

$$\beta_1 = \frac{\mathrm{Cov}(Y_i, X_i)}{\mathrm{Var}(X_i)}$$

- In the multivariate case, β_j is:

$$\beta_j = \frac{\mathrm{Cov}(Y_i, \tilde{X}_{ij})}{\mathrm{Var}(\tilde{X}_{ij})}$$

  where $\tilde{X}_{ij}$ is the residual from the regression of X_{ij} on all other covariates.
- The multiple regression coefficient $\hat{\beta}_j$ represents the additional contribution of x_j on y, after x_j has been adjusted for $x_o, x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$
- What happens when x_j is highly correlated with some of the other x_k's?

Regression in Causal Analysis

- Imagine we are analyzing a *randomized* experiment with a regression using the following model:

$$Y_i = \alpha + \beta_1 \cdot T_i + \mathbf{X}_i^T \cdot \beta_2 + \epsilon_i$$

  where T_i is an indicator variable for treatment status and $\mathbf{X}_i$ is a vector of *pre-treatment characteristics*
- Under this model, what is random?

Regression in Causal Analysis

- Imagine we are analyzing a *randomized* experiment with a regression using the following model:

$$Y_i = \alpha + \beta_1 \cdot T_i + \mathbf{X}_i^T \cdot \beta_2 + \epsilon_i$$

  where T_i is an indicator variable for treatment status and $\mathbf{X}_i$ is a vector of *pre-treatment characteristics*
- Under this model, what is random?
- How do we interpret the coefficients on $\mathbf{X}_i$?

Regression in Causal Analysis

- Imagine we are analyzing a *randomized* experiment with a regression using the following model:

$$Y_i = \alpha + \beta_1 \cdot T_i + \mathbf{X}_i^T \cdot \beta_2 + \epsilon_i$$

  where T_i is an indicator variable for treatment status and $\mathbf{X}_i$ is a vector of *pre-treatment characteristics*
- Under this model, what is random?
- How do we interpret the coefficients on $\mathbf{X}_i$?
- How do we interpret the coefficient β_1?

Regression in an Observational Study