36-617: Applied Linear Models

Multivariate Regression

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Reading, HW, Quiz

- Reading
  - This week: Sheather Ch 5 (Skip Ch 4 for now)
  - For next week: Sheather Ch 6

- HW 03 due tonight 1159pm
- HW 04 will be on Canvas later today – Due next Monday

- Quiz on Ch 5 in class – today!

Outline

- Matrix Form of Multiple Regression Model
- Multivariate Normal Distribution
- ML/LS Estimates
- Two Interpretations
- Distributional Properties
- SS Decompositions and F Statistics
- Some Comments

Matrix Form of Multiple Regression Model

- Let $X_i = (x_{i0}, \ldots, x_{ip})$ and $\beta = (\beta_0, \ldots, \beta_p)^T$; then

$$\begin{aligned} y_i &= \beta_0 X_{i0} + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i \\ &= X_i \beta + \epsilon_i \end{aligned}$$

- If we also stack $Y = (y_1, \ldots, y_n)^T$, $X = (X_1^T, \ldots, X_n^T)^T$, and $\epsilon = (\epsilon_1, \ldots \epsilon_n)^T$, we can write

$$Y_{n\times 1} = X_{n\times k}\beta_{k\times 1} + \epsilon_{n\times 1} \quad (k = p+1)$$

Matrix Form of Multiple Regression Model

$$Y = X\beta + \epsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ x_{20} & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Usually $x_{i0} \equiv 1$, so we get

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Multivariate Normal Distribution

- In the model

$$y_i = X_i\beta + \epsilon_i, i = 1, \ldots, n$$

  it is usual to assume $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

- Recall
  - Y ~ N(0,1) iff $f(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$
  - Y ~ N(μ,σ^2) iff $\frac{y-\mu}{\sigma} \sim N(0,1)$

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Multivariate Normal Distribution

- $Y = (Y_1, \ldots, Y_n)^T \sim N(0, I) = N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}\right)$

$$f(y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-y_i^2/2}$$

- $Y \sim N(\mu, \Sigma)$ iff

$$\Sigma^{-1/2}(Y - \mu) \sim N(0, I)$$

…and some ugly formula for $f(y_1, \ldots, y_n)$…

Multivariate Normal Distribution

- When Y ~ N(μ,$\sum$), then

$$\mu = (\mu_1, \mu_2, \ldots, \mu_n)^T = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

is the *mean vector*:

$$E[y_i] = \mu_i, \ i = 1, \ldots, n$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

is the *variance-covariance matrix*:

$$\begin{aligned} \text{Var}(y_i) &= \sigma_i^2, \ i = 1, \ldots, n \\ \text{Cov}(y_i, y_j) &= \sigma_{ij}, \ i, j = 1, \ldots, n \end{aligned}$$

Many Equivalent Model Formulations

- $Y \sim N(X\beta, \sigma^2 I)$, i.e.:
  - $E[y_i] = X_i\beta$, i = 1, …,n
  - $Var(y_i) = \sigma^2$, i=1, …, n
  - $Cov(y_i,y_j) = 0, \forall\ i \neq j$
- We could also write
  - $Y = X\beta + \epsilon, \epsilon \sim N(0,\sigma^2 I)$
  - $y_i = X_i\beta + \epsilon_i, \epsilon_i \sim N(0,\sigma^2)$, iid
  - $y_i = \beta_1 X_{i1} + \ldots + \beta_k X_{iK} + \epsilon, \epsilon_i \sim N(0,\sigma^2)$, iid

ML/LS Estimates

- $Y_i = X_i\beta + \epsilon_i$, $\epsilon_i \sim N(0,\sigma^2)$
- So $Var(Y_i) = E[(Y_i - X_i\beta)^2] = E[\epsilon_i^2] = Var(\epsilon_i) = \sigma^2$
- Then we can estimate (MoM!):

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - X_i\beta)^2$$

- Fitting the model is basically just finding values β to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - X_i\beta)^2 = \frac{1}{n}(Y - X\beta)^T(Y - X\beta)$$

- It turns out that $\hat{\beta} = (X^TX)^{-1}X^Ty$

ML/LS Estimates

$$y = X\beta + \epsilon,\ \ \epsilon \sim N(0, \sigma^2 I)$$

- $\hat{\beta} = (X^T X)^{-1} X^T y$
- $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$
- The "residual SD" is the square root of

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} (y_i - X_i \hat{\beta})^2 = \frac{1}{n-k} (y - X\hat{\beta})^T (y - X\hat{\beta})$$

- We will see below that

$$\begin{aligned} \mathrm{Var}(\hat{\beta}) &= (X^T X)^{-1} \sigma^2 \\ \mathrm{Var}(\hat{y}) &= X(X^T X)^{-1} X^T \sigma^2 \end{aligned}$$

Recall that k=p+1

ML/LS Estimates - Example

```
> str(kidiq)
> y <- cbind(kidiq$kid.score)
> X <- cbind(1, with(kidiq,
   cbind(mom.hs,mom.iq)))
> dim(X)
[1] 434   3
> n <- dim(X)[1]
> k <- dim(X)[2]
> V <- solve(t(X) %*% X)
> beta.hat <- V %*% t(X) %*% y
> res.var <- t(y - X%*%beta.hat)
   %*% (y - X%*%beta.hat) / (n-k)
> res.sd <- sqrt(res.var)
>
> var.beta <- V * c(res.var)
> beta0.sd <- sqrt(var.beta[1,1])
> beta1.sd <- sqrt(var.beta[2,2])
> beta2.sd <- sqrt(var.beta[3,3])
```

```
> round(cbind(beta.hat,
   c(beta0.sd, beta1.sd,
   beta2.sd)), 2)
          [,1] [,2]
       25.73 5.88
mom.hs  5.95 2.21
mom.iq  0.56 0.06
> round(res.sd,2)
      [,1]
[1,] 18.14
```

- `summary(lm(kid.score ~ mom.hs + mom.iq, data=kidiq))`

```
             coef.est  coef.se
(Intercept) 25.73     5.88
mom.hs       5.95     2.21
mom.iq       0.56     0.06
---
n = 434, k = 3
res sd = 18.14, R-Squared = 0.21
```

estimates-by-hand.r

Two Interpretations

- In the model $y_i = \beta_o X_{i0} + \ldots + \beta_p X_{ip} + \epsilon_i$, β_j is the *change in y for a unit change in X_j, holding the other X's fixed*
  - Since $\hat{\beta}_j$ estimates β_j, $\hat{\beta}_j$ inherits this interpretation also.
- When we look at "added variable plots", we will see that $\hat{\beta}_j$ measures *the variation in y, left after controlling for the other X's, that is uniquely attributable to X_j.*

Distributional Properties: $\hat{\beta}$

Fact: $Y \sim N(\mu, \Sigma) \Rightarrow AY \sim N(A\mu, A\Sigma A^T)$

$$
\begin{aligned}
y &\sim N(X\beta, \sigma^2 I) \\
\hat{\beta} &= (X^T X)^{-1} X^T y \\
E[\hat{\beta}] &= E[(X^T X)^{-1} X^T y] = (X^T X)^{-1} X^T E[y] \\
&= (X^T X)^{-1} X^T X \beta = \beta \\
\mathsf{Var}(\hat{\beta}) &= \mathsf{Var}((X^T X)^{-1} X^T y) \\
&= (X^T X)^{-1} X^T \mathsf{Var}(y) X (X^T X)^{-1} \\
&= (X^T X)^{-1} X^T X (X^T X)^{-1} \sigma^2 = (X^T X)^{-1} \sigma^2 \\
\Rightarrow \hat{\beta} &\sim N(\beta, (X^T X)^{-1} \sigma^2)
\end{aligned}
$$

Hat Matrix H, & Distribution of $\hat{y}$ and $\hat{e}$

$$\hat{y} = X\hat{\beta} = X\left[(X^TX)^{-1}X^Ty\right] = \underbrace{\left[X(X^TX)^{-1}X^T\right]}_{\text{Hat matrix, H}}y = Hy$$

$$HX = X(X^TX)^{-1}X^TX = X$$

$$\Rightarrow \forall\beta^*, HX\beta^* = X\beta^*$$

$$H^T = H \quad \text{and} \quad (I-H)^T = (I-H)$$

$$HH = H \quad \text{and} \quad (I-H)(I-H) = (I-H)$$

$$E[\hat{y}] = E[Hy] = HE[y] = HX\beta = X\beta$$

$$\text{Var}(\hat{y}) = \text{Var}(Hy) = H\text{Var}(y)H^T = HH\sigma^2 = H\sigma^2$$

$$\Rightarrow \hat{y} \equiv Hy \sim N(X\beta, H\sigma^2)$$

Similarly, $\hat{e} = y - \hat{y} \equiv (I-H)y \sim N(0, (I-H)\sigma^2)$

SS Decompositions and F Statistics

Fact: $y^T A y + y^T B y = (y^T A + y^T B) y = y^T (A + B) y$

$$
\begin{aligned}
SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = (y - \bar{y})^T (y - \bar{y}) \\
&= y^T (I - H_1)^T (I - H_1) y = y^T (I - H_1) y \\
SS_{reg} &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = (\hat{y} - \bar{y})^T (\hat{y} - \bar{y}) = y^T (H - H_1) y \\
RSS &= \sum_{i=1}^{n} (y_i - \hat{y})^2 = (y - \hat{y})^T (y - \hat{y}) = y^T (I - H) y
\end{aligned}
$$

$$
\Rightarrow SS_{reg} + RSS = y^T (H - H_1) y + y^T (I - H) y = y^T (I - H_1) y = SST
$$

- *Cochran's Theorem* implies: SS_{reg}/σ^2 and RSS/σ^2 are indep. χ^2's under $H_0 : \beta_1 = \cdots = \beta_p = 0$
- The df for each χ^2 will be the rank, or equivalently the trace, of each defining matrix. Using $tr(AB) = tr(BA)$: $tr(H) = p + 1$, $tr(H_1) = 1$, $tr(I) = n$, so

$$
df(SS_{reg}/\sigma^2) = p,\ df(RSS/\sigma^2) = n - p - 1,\ df(SST/\sigma^2) = n - 1
$$

(H_1 = hat matrix for $y = \beta_0 + \varepsilon$)

SS Decompositions and F Statistics

- The foregoing lead to the traditional Analysis of Variance Table

Source of variation	Degrees of freedom (df)	Sums of squares (SS)	Mean square (MS)	F
Regression	p	$SSreg$	$SSreg/p$	$F = \frac{SSreg/p}{RSS/(n-p-1)}$
Residual	$n-p-1$	RSS	$RSS/(n-p-1)$	
Total	$n-1$	SST		

- As before we can define "multiple R^2":

$$R^2 = \frac{SS_{reg}}{SST} = 1 - \frac{RSS}{SST} \quad (\; = \; \mathrm{Corr}\,(y, \hat{y})^2 \;)$$

- "Adjusted R^2": mean-squares instead of sums of squares, to account for capitalization on chance

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$$

SS Decompositions and F Statistics

Since $SST = y^T(I - H_1)y = RSS_{H_1}$, the residual sum-of-square from the smallest model (intercept-only), the F statistic from the Anova table can be written as

$$F = \frac{SS_{reg}/p}{RSS/n-p-1} = \frac{(RSS_{H_1} - RSS_H)/(df_{H_1} - df_H)}{RSS_H/df_H} \qquad (*)$$

This idea, and the sum-of-square calculations we did earlier, can be generalized so that, if H_{full} and $H_{reduced}$ are hat matrices from a "full" model, and from a "reduced" model obtained by linear restrictions on the "full" model, then the ***partial F statistic***

$$F = \frac{(RSS_{H_{reduced}} - RSS_{H_{full}})/(df_{H_{reduced}} - df_{H_{full}})}{RSS_{H_{full}}/df_{H_{full}}}$$

will have an F distribution under the null hypothesis that the linear restrictions are true.

(*) The df throughout are *residual* df, that is, tr(I-H)

Some Comments

- It's good to know the "canonical" theory of the linear model and the Analysis of Variance table
  - Distribution assumptions and multiple testing matters
  - We will more fully discuss later in the course
- The "linear restrictions" for the partial F statistic usually amount to just setting some β's = 0 . This is especially useful when a regressor is categorical, since a categorical X is recoded as a set of dummy variables, one for each level of X
- The partial F test brings us into "variable selection"
  - We will more fully discuss variable selection later as well!

Summary

- Matrix Form of Multiple Regression Model
- Multivariate Normal Distribution
- ML/LS Estimates
- Two Interpretations
- Distributional Properties
- SS Decompositions and F Statistics
- Some Comments