Extending SLR to MLR through the Matrix Form

Overview of Multiple Linear Regression (MLR)

Data for Multiple Regression

- Y_i is the response variable (as usual)
- $X_{i,1}, X_{i,2}, \ldots, X_{i,p-1}$ are the $p-1$ explanatory variables for cases $i = 1$ to n.
- Example – In Homework #1 you modeled GPA as a function of entrance exam score. We could also consider an aptitude test and high school GPA as potential predictors. With the entrance exam score, this would be 3 variables, so $p = 4$.
- *Potential problem to remember!!!* These predictor variables are probably correlated with each other.

The Multiple Regression Model in Scalar Form
(MLR: multiple linear regression)

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \text{ for } i = 1, 2, \ldots, n$$

where

- Y_i is the value of the response variable for the ith case.
- $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ (exactly as before!)
- β_0 is the intercept (think multidimensionally and look at the equation).
- $\beta_1, \beta_2, \ldots, \beta_{p-1}$ are the regression coefficients for the explanatory variables.
- $X_{i,k}$ is the value of the kth explanatory variable for the ith case.

Special Cases

- *Polynomial model*

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_{p-1} X_i^{p-1} + \varepsilon_i$$

- *Interactions* between explanatory variables are expressed as a product of the X's:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,1} X_{i,2} + \varepsilon_i$$

- *ANOVA Models* with discrete predictors can be encoded by defining the X's as *indicator* or *dummy variables* where $X_{i,k} = 1$ if case i belongs to the k-th group, and $X = 0$ otherwise.

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,1} X_{i,2} + \varepsilon_i$$

- *Linear model vs Nonlinear model (not covered in the course)*

$$e.g.\ Y_i = \beta_0 \exp(\beta_1 X_i) + \varepsilon_i$$

Multiple Regression Model in Matrix Form

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}_{n\times n})$$

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$$

Design Matrix x :

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} \\ & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p-1} \end{bmatrix}$$

Matrix Forms for the MLR are Identical to the SLR

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \dots \\ Y_n \end{bmatrix} \sim \text{Normal}\left(\begin{bmatrix} X_1\beta \\ X_2\beta \\ X_3\beta \\ \dots \\ X_n\beta \end{bmatrix}, \begin{bmatrix} \sigma^2 \ 0 \ 0 \ \dots 0 \\ 0 \ \sigma^2 \ 0 \ \dots 0 \\ 0 \ 0 \ \sigma^2 \ \dots 0 \\ \dots \\ 0 \ 0 \ 0 \ \dots \sigma^2 \end{bmatrix} \right)$$

$$\Sigma\{Y\}_{nXn} = \Sigma\{\varepsilon\}_{nXn} = \sigma^2 I$$

- $Y = X\beta + \varepsilon = X\beta + I\varepsilon$, and
- No assumption violation

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

$$b = (X^tX)^{-1}X^tY$$

$$\Sigma\{b\}_{pXp} = \sigma^2(X'X)^{-1} = \text{MSE}\,(X'X)^{-1}$$

$$\hat{Y}_h = \mathbf{X}_h'\mathbf{b}$$

$$\Sigma\{\hat{Y}_h\} = X_h'\Sigma\{b\}X_h$$

Matrix Forms for the Residuals

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

$$\Sigma\{\mathrm{e}\}_{\mathrm{nXn}} = (\mathrm{I} - \mathrm{H})\Sigma\{\mathrm{Y}\}(\mathrm{I} - \mathrm{H})'$$
$$= (\mathrm{I} - \mathrm{H})\sigma^2\mathrm{I}(\mathrm{I} - \mathrm{H})' = \sigma^2(\mathrm{I} - \mathrm{H})(\mathrm{I} - \mathrm{H})'$$
$$= \sigma^2(\mathrm{I} - \mathrm{H}) = \mathrm{MSE}(\mathrm{I} - \mathrm{H})$$

$\sigma^2(e_{\mathrm{i}}) = \mathrm{MSE}(1 - \mathrm{h}_{\mathrm{ii}})$, where h_{ii} is the i-th diagonal element of H.

$h_{ii} = X_i'(X'X)^{-1}X_i \quad and \;\; X_i' = \left(1 \;\; X_{i1}, \ldots. X_{ip-1}\right)$ is taken from the i-th data point.

The covariance $\sigma\left(e_{\mathrm{i}}, \mathrm{e}_j\right) = \mathrm{MSE}(-\mathrm{h}_{\mathrm{ii}})$ is usually not 0, but we can ignore this with a reasonably large n.

In the flavor example,

$$\mathrm{MSE}(\boldsymbol{I} - \boldsymbol{H}) = 0.0493 \quad \begin{matrix} 0.4 & -0.4 & -0.2 & 0.0 & 0.2 \\ -0.4 & 0.7 & -0.2 & -0.1 & 0.0 \\ -0.2 & -0.2 & 0.8 & -0.2 & -0.2 \\ 0.0 & -0.1 & -0.2 & 0.7 & -0.4 \\ 0.2 & 0.0 & -0.2 & -0.4 & 0.4 \end{matrix}$$

The ANOVA Table is Identical to the SLR

Source	df	SS	MSE	F
Model	$df_M = p - 1$	SSM	MSM	$\frac{MSM}{MSE}$
Error	$df_E = n - p$	SSE	MSE	
Total	$df_T = n - 1$	SST		

The Global F test or the Significant Test

$H_0: \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$ (β_0 is not in this list!)

$H_A: \beta_k \neq 0$, for at least one $k = 1, \ldots, p-1$

$F^* = MSM/MSE$

Under H_0, $F^* \sim F_{p-1,n-p}$.

We test it the usual way (reject H_0 if $p \leq \alpha$).

Interpreting the p-value of the Global F-test

If the p-value for the F-test . . .

- is $> \alpha$, we lack evidence to conclude that *any* of our explanatory variables can help to predict or explain the response variable using a linear regression model.
- is $\leq \alpha$, *one or more* of the explanatory variables in our model *is* potentially useful for predicting the response in a linear model (but F does not say which ones).

Coefficient of Multiple Determination, R^2

As in SLR, R^2 is the proportion of variation in the response explained by the model.
R^2 and Fs are related.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad F = \frac{\frac{R^2}{p-1}}{\frac{1-R^2}{n-p}}$$

Note that "explained by the model" means "explained by these predictors using this specific regression equation," *NOT* just "explained by these predictor variables."

Finally, a large value of R^2 does not necessarily imply that the fitted model is a useful one.

Adjusted Coefficient of Determination, R^2_{adj}

It is sometimes suggested that a modified measure be used that adjusts for the number of X variables in the model.

$$\mathrm{R}^2_{\mathrm{adj}} = 1 - \frac{MSE}{MST} = 1 - \left(\frac{n-1}{n-p}\right) SSE/SST$$

- $1/df_E = 1/(n-p)$ increases as a hyperbolic function of p.
- Increasing p by 1 *always* makes SSE smaller, but not always by the same amount.
- If the decline in SSE is large enough to cancel the increase in $1/df_E$, then MSE will get smaller, so R^2 will get bigger.
- If the fit does not improve sufficiently to overcome $1/df_E$, then R^2_{Adj} will remain the same, or might even get smaller!

The T test on the Individual Regression Coefficients (Parameters)

As usual, the CI for β_k is,

$$b_k \pm t_c s_{\{b_k\}}, \text{ where } t_c = t_{n-p}(1-\alpha/2)$$

We know that $\boldsymbol{b} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1})$

We estimate the variance-covariance matrix for the parameter vector as,

$$\Sigma\{b\}_{pXp} = MSE\,(X^tX)^{-1}$$
$$= \left(\frac{1}{n-p}\right) Y'(I-H)Y(X^tX)^{-1}$$

For an individual coefficient β_k, where $k = (0, \ldots p-1)$,
$s^2_{\{b_k\}}$ is the $(k+1)$-th diagonal element of the variance-covariance matrix $\Sigma\{b\}_{pXp}$

The T test on the Individual Regression Coefficients (Parameters)

- The hypothesis test is defined as $\mathrm{Ho}: \beta_k = \beta_k^*$. By default, $\beta_k^* = 0$ and two-sided.
- This tests the significance of this β_k, given all other βs in the model.

For example：Ho: $\beta_3 = 0 \mid \beta_1, \beta_2, \beta_4$ in the model Ha: $\beta_3 \neq 0 \mid \beta_1, \beta_2, \beta_4$ in the model

- The test statistic is $t_s = \frac{b_k}{\mathrm{s}\{b_k\}} \sim \mathrm{t}(\mathrm{n} - p)$

$$\mathrm{ts}^2 \neq Fs = \frac{MSR}{MSE}$$

- The result of the Significant test of beta could be misleading when the impact of X_k overlaps with other predictors.

The Dwaine Studios example: The Dwaine operates studios that specialize in portraits of children. The company is considering whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X1) and the per capita disposable personal income in the community (X2). Data is Dwaine.csv, n=21, p=3

X1	X2	Y
68.5	16.7	174.4
45.2	16.8	164.4
91.3	18.2	244.2
47.8	16.3	154.6
46.9	17.3	181.6
66.1	18.2	207.5
49.5	15.9	152.8
52	17.2	163.2
48.9	16.6	145.4
38.4	16	137.2
87.9	18.3	241.9
72.8	17.1	191.1
88.4	17.4	232
42.9	15.8	145.3
52.5	17.8	161.1
85.7	18.4	209.7
41.3	16.5	146.4
51.7	16.3	144
89.6	18.1	232.6
82.7	19.1	224.1
52.3	16	166.5

The Dwaine Studios example: The Dwaine operates studios that specialize in portraits of children. The company is considering whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X1) and the per capita disposable personal income in the community (X2). Data is Dwaine.csv, n=21, p=3

1). Predict the mean Y when X1=65.4 and X2=17.6

$$\hat{Y} = \quad b_0 + b_1 X_1 + b_2 X_2$$
$$= 191$$

2). Find SSE, df_E, SSR, df_R, MSE, MSR, R^2, R^2_{adj}

SSE= 2180.9 dfE= 18 MSE=121.2

SSR=23371.8+643.5 =24015.3, dfR = p-1=2

MSR= 24015.3/2 =12007.6

SST=SSR+SSE= 26195.3

$R^2 = \quad SSR/SS = 0.9167$

$R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right) SSE/SST$ = 0.9075

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.8571    60.0170  -1.147   0.2663
x1            1.4546     0.2118   6.868    2e-06 ***
x2            9.3655     4.0640   2.305   0.0333 *

Residual standard error: 11.01 on 18 degrees of freedom
Multiple R-squared:  0.9167,    Adjusted R-squared:  0.9075
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10
```

```
Response: y
          Df  Sum Sq Mean Sq  F value    Pr(>F)
x1         1 23371.8 23371.8 192.8962 4.64e-11 ***
x2         1   643.5   643.5   5.3108  0.03332 *
Residuals 18  2180.9   121.2
```

```
dwa.mod<-lm(y~x1+x2, dwaine)
summary(dwa.mod)
anova(dwa.mod)
```

The Dwaine Studios example: The Dwaine operates studios that specialize in portraits of children. The company is considering whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X1) and the per capita disposable personal income in the community (X2). Data is Dwaine.csv, n=21, p=3

3). Find the estimated variance-covariance matrix for the parameter $\Sigma^2\{b\} = MSE(X'X)^{-1}$

$s^2\{b_1\}$ =0.04485

$s^2\{b_2\} = 16.5158$

	(Intercept)	x1	x2
(Intercept)	3602.03467	8.74593958	-241.4229923
x1	8.74594	0.04485151	-0.6724426
x2	-241.42299	-0.67244260	16.5157558

$\boldsymbol{Cov(b_1, b_2) = -0.67}$

`vcov(dwa.mod)`

4). Test whether sales are related to the target population and per capita disposable income.

$H_0: \beta_1 = 0 \; and \; \beta_2 = 0$,
H_a: not both $\beta_1 and \; \beta_2 \; equal \; zero$

$F^* = \; MSR/MSE \; = 99.1$

For $\alpha = 0.05$, we require $F(0.95; 2{,}18) = 3.55$.
The sales are related to *(at least one or both)* target population and per capita disposable income

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.8571    60.0170  -1.147   0.2663
x1            1.4546     0.2118   6.868    2e-06 ***
x2            9.3655     4.0640   2.305   0.0333 *

Residual standard error: 11.01 on 18 degrees of freedom
Multiple R-squared:  0.9167,    Adjusted R-squared:  0.9075
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10

Response: y
          Df  Sum Sq Mean Sq  F value   Pr(>F)
x1         1 23371.8 23371.8 192.8962 4.64e-11 ***
x2         1   643.5   643.5   5.3108  0.03332 *
Residuals 18  2180.9   121.2
```

The Dwaine Studios example: The Dwaine operates studios that specialize in portraits of children. The company is considering whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X1) and the per capita disposable personal income in the community (X2). Data is Dwaine.csv, n=21, p=3

5) Find the 95% *individual* CI for each of the parameter.

95% CI for β_1: $b_1 \pm t\ \ s\{b1\}$=$1.455 \pm 2.101(0.2118)$
= (1.01, 1.9}

95% CI for β_2: $b_2 \pm t\ \ s\{b2\}$=$9.366 \pm 2.101(4.064)$
= (0.83, 17.9}

```
            (Intercept)          x1           x2
(Intercept)  3602.03467  8.74593958 -241.4229923
x1              8.74594  0.04485151   -0.6724426
x2           -241.42299 -0.67244260   16.5157558
                   2.5 %    97.5 %
(Intercept) -194.9480130 57.233867
x1             1.0096226  1.899497
x2             0.8274411 17.903560
confint(dwa.mod,level=0.95)
```

6) Find the 95% simultaneous Bonferroni CI for the parameter (β_1 *and* β_2), g=2

$$\mathrm{B} = t\left(1-\frac{\alpha}{2g}, dfE\right) = \mathrm{t}(1-0.9875, 18) = 2.445$$

95% simultaneous Bonferroni CI for β_1 :

$$b_1 \pm t(1-\alpha/2g, df)s\{b1\}$$
$$= 1.45 \pm 2.445(0.2118) = (0.932, 1.968)$$

95% simultaneous Bonferroni CI for β_2 :

$$b_2 \pm t(1-\alpha/2g, df)s\{b2\}$$
$$= 9.37 \pm 2.445(\ 4.064) = (-0.566, 19.306)$$