Multiple Regression

- Multiple Regression ANOVA
- Standardized Regression
- Multi-collinearity
- Polynomial Regression
- Interaction Regression Model
- Constrained Regression
Extra sum of squares: decompose SSR

Consider the following regression model with two covariates,

\[ Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon, \]

note the ANOVA as

\[ SSTO = SSR(X_1, X_2) + SSE(X_1, X_2), \]

for regression model with just \( X_1 \)

\[ Y = \beta_0 + X_1\beta_1 + \epsilon, \]

the ANOVA is

\[ SSTO = SSR(X_1) + SSE(X_1). \]
Obviously

\[ SSE(X_1, X_2) \leq SSE(X_1), \]
or equivalently

\[ SSR(X_1, X_2) \geq SSR(X_1). \]

Define the extra (regression) sum of squares as

\[ SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1) \]
\[ = SSE(X_1) - SSE(X_1, X_2). \]  \(1\)

This increase (reduction) in the regression (error) sum of squares is the result of adding \( X_2 \) to the regression model when \( X_1 \) is already included in the model.

Notice that generally

\[ SSR(X_2|X_1) \neq SSR(X_1|X_2), \quad SSR(X_1) \neq SSR(X_2). \]
We can similarly define general extra sum of squares

\[ SSR(X_S|X_F) = SSR(X_S, X_F) - SSR(X_F), \]  

(2)

where \( X_S, X_F \) are two sets of covariates, e.g.

\[ SSR(X_2, X_3|X_1) = SSR(X_1, X_2, X_3) - SSR(X_1), \]

\[ SSR(X_3|X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2). \]

By definition we have

\[ SSR(X_2, X_3|X_1) = SSR(X_2|X_1) + SSR(X_3|X_1, X_2), \]

and

\[ SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2) \]
Extra sum of squares for testing regression coefficients

Consider the following general multiple regression model

\[ Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon, \]

to test a single regression coefficient \( \beta_k \)

\[ H_0 : \beta_k = 0 \ vs \ H_a : \beta_k \neq 0, \]  

(3)

we can use the \( t \)-test

\[ t^* = \frac{b_k}{s(b_k)} \sim t(n - p - 1) \text{ under } H_0. \]
Equivalently we can use the following general linear test approach: compare the reduced model (R)

\[ \mathbb{E}(Y) = \beta_0 + \sum_{j \neq k} X_j \beta_j, \]

to the full model (F)

\[ \mathbb{E}(Y) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j, \]

if \( \beta_k = 0 \), the two \( SSE \) should be very close. We can use the following \textit{partial F test} statistic

\[
F^* = \frac{SSE(R) - SSE(F)}{1} \div \frac{SSE(F)}{n - p - 1} = \frac{SSR(X_k|X_{-k})}{SSE(X_1, \cdots, X_p)/(n - p - 1)} \sim F(1, n - p - 1) \text{ under } H_0.
\]
We can show that

\[ F^* = t^*^2, \]

so the two tests are equivalent.

The idea can be carried over to test for several regression coefficients being zero

\[ H_0 : \beta_i = 0, \forall i \in S, \quad H_a : \exists i \in S, \beta_i \neq 0, \]

(4)

where \( S \) is a collection of integers, e.g. for \( S = 1, 2, 3 \), we’re testing \( \beta_1, \beta_2 \) and \( \beta_3 \) being zero simultaneously. The \textit{partial F test} statistic being

\[
F^* = \frac{SSR(X_S|X-S)/\#(S)}{SSE(X_1, \ldots, X_p)/(n-p-1)} \sim F(\#(S), n-p-1) \text{ under } H_0,
\]

where \( \#(S) \) equals the number of elements in set \( S \).

We can also write the regression sum of squares for the full model in an
extra sum of squares notation as

$$SSR = SSR(X_1, \ldots, X_p|1),$$

which is the extra variation explained by adding all covariates in the model. And hence we have the following F test

$$F = \frac{SSR/p}{SSE/(n - p - 1)}$$

for testing all $\beta_k$ being zero.

We can also test for some linear equations of some regression coefficients, e.g.

$$H_0 : \beta_1 = \beta_2 \text{ vs } H_a : \beta_1 \neq \beta_2,$$  \hspace{1cm} (5)
we can compare the SSE of the following two regression models

\[ E(Y) = \beta_0 + \sum_{j} X_j \beta_j \text{ vs } E(Y) = \beta_0 + \beta_+(X_1 + X_2) + \sum_{j=3}^{p} X_j \beta_j, \]

note that we can equivalently write the full model as

\[ E(Y) = \beta_0 + \beta_+ (X_1 + X_2) + \beta_- (X_1 - X_2) + \sum_{j=3}^{p} X_j \beta_j, \]

so the testing transforms to the \textit{extra sum of squares} framework.

When testing

\[ H_0 : \beta_1 = 2, \beta_2 = 4, \text{ vs } H_a : \text{ not both equations in } H_0 \text{ are true} \] (6)
we can compare the following two regression models

\[ E(Y) = \beta_0 + \sum_j X_j \beta_j \text{ vs } E(Y - 2X_1 - 4X_2) = \beta_0 + \sum_{j=3}^p X_j \beta_j , \]

Viewed from the projection perspective, for all previous tests, the test statistics are comparing two orthogonal projections of the model errors under \( H_0 \).
The *coefficients of partial determination* measures the marginal contribution of a variable when all other variables are already included in the model. E.g.

\[
\begin{align*}
 r^2_{Y1.2} &= \frac{SSR(X_1|X_2)}{SSE(X_2)}, \\
 r^2_{Y2.1} &= \frac{SSR(X_2|X_1)}{SSE(X_1)},
\end{align*}
\]

where \( r^2_{Y1.2} \) measures the proportionate reduction in the variation in \( Y \) remaining after \( X_2 \) is included in the model that is gained by also including \( X_1 \) in the model.
Generally we can define

\[ r^2_{Y_{k.S}} = \frac{SSR(X_k|X_S)}{SSE(X_S)}, \]  

(7)

where \( S \) are a set of indicies for the variables and \( k \) an integer, e.g.

\[ r^2_{Y_{3.12}} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)}. \]

We can extend the definition to measure the relative contribution for a set of variables

\[ r^2_{Y_{K.S}} = \frac{SSR(X_K|X_S)}{SSE(X_S)}, \]  

(8)
where $K, S$ are two sets of indices, e.g.

$$r_{Y34.12}^2 = \frac{SSR(X_3, X_4|X_1, X_2)}{SSE(X_1, X_2)}.$$

And we can explain the coefficient of partial determination using coefficient of determination. Just notice that $SSE(X_S)$ is the SSTO for the residuals of regressing $Y$ on $X_S$, and $SSR(X_K|X_S)$ is the SSR for regressing the $Y$ residuals on the set of variables which are the residuals of regressing $X_K$ on $X_S$. 
Coefficient of Partial Correlation

The square root of a coefficient of partial determination is called a coefficient of partial correlation, where the sign is the same as the corresponding regression coefficient in the model.

\[ r_{Yk.S} = \sqrt{r_{YK.S}^2 \times \text{sgn}(b_k)}. \]  

(9)
Correlation Transformation

Rounding errors tend to occur when the $X$ variables have substantially different magnitudes so that the entries in the $X'X$ matrix cover a wide range, say, from $10^{-2}$ to $10^8$.

Another difficulty with the multiple regression is that ordinarily regression coefficients cannot be compared because of different units used by different covariates.

Correlation transformation can be used to re-parameterize the regression model into the standardized regression model. It is a simple modification of the usual variable standardization, which involves centering and scaling.
variables

\[ \frac{Y_i - \bar{Y}}{s(Y)}, \quad s(Y) = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}}, \]

\[ \frac{X_{ik} - \bar{X}_k}{s(X_k)}, \quad s(X_k) = \sqrt{\frac{\sum_{i=1}^{n} (X_{ik} - \bar{X}_k)^2}{n-1}}, \]

The correlation transformation is the following

\[ Y_i' = \frac{1}{\sqrt{n-1}} \left( \frac{Y_i - \bar{Y}}{s(Y)} \right), \]

\[ X_{ik}' = \frac{1}{\sqrt{n-1}} \left( \frac{X_{ik} - \bar{X}_k}{s(X_k)} \right), \quad (10) \]

where the multiplied constant will make

\[ \sum_{i=1}^{n} Y_i'^2 = \sum_{i=1}^{n} X_{ik}'^2 = 1. \]
Standardized Regression Model

The *standardized regression model* refers to regressing the transformed $Y'$ on $X'_k$

$$Y'_i = X'_{i1}\beta_1' + \cdots + X'_{ip}\beta_p' + \epsilon'_i, \quad (11)$$

and it is easy to verify that

$$\beta_k = \frac{s(Y)}{s(X_k)}\beta_k', \quad k = 1, \cdots, p$$

$$\beta_0 = \bar{Y} - \bar{X}_1\beta_1 - \cdots - \bar{X}_p\beta_p. \quad (12)$$

It can be shown that for the standardized regression model,

$$X'X = r_{XX},$$
where \( r_{XX} \) is the correlation matrix of the \( X \) variables

\[
r_{XX} = \begin{bmatrix}
1 & r_{12} & \cdots & r_{1p} \\
 r_{21} & 1 & \cdots & r_{2p} \\
 \vdots & \vdots & & \vdots \\
 r_{p1} & r_{p2} & \cdots & r_{pp}
\end{bmatrix}
\]

where

\[
r_{ij} = \frac{\sum_{k=1}^{n}(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)}{\sqrt{\sum (X_{ki} - \bar{X}_i)^2 \sum (X_{kj} - \bar{X}_j)^2}} \in [-1, 1].
\]

So all the elements of \( X'X \) are between -1 and 1, and thus are of the same order of magnitude, which can be of great help in controlling roundoff errors when inverting \( X'X \).
It is easy to show that for the standardized regression model

\[ X'Y = r_{YX} = \begin{bmatrix} r_{Y_1} \\ r_{Y_2} \\ \vdots \\ r_{Y_p} \end{bmatrix} \]

where

\[ r_{Y_j} = \frac{\sum_{k=1}^{n} (Y_k - \bar{Y})(X_{kj} - \bar{X}_j)}{\sqrt{\sum (Y_k - \bar{Y})^2 \sum (X_{kj} - \bar{X}_j)^2}}. \]

So the standardized regression coefficients are

\[ b = r_{XX}^{-1} r_{YX}. \]
E.g. for $p = 2$ we have

$$r_{XX} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}, \quad r_{YX} = \begin{bmatrix} r_{Y1} \\ r_{Y2} \end{bmatrix},$$

$$b = r_{XX}^{-1} r_{YX} = \frac{1}{1 - r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix} \begin{bmatrix} r_{Y1} \\ r_{Y2} \end{bmatrix},$$

so

$$b'_1 = \frac{r_{Y1} - r_{12}r_{Y2}}{1 - r_{12}^2}, \quad b'_2 = \frac{r_{Y2} - r_{12}r_{Y1}}{1 - r_{12}^2}.$$
Uncorrelated Predictors

When predictor variables are correlated among themselves, i.e. \( \exists i, j, r_{ij} \neq 0 \), intercorrelation or multicollinearity is said to exist. When \( r_{ij} = 0, \forall i, j \), the predictors are said to be uncorrelated.

Notice that the least squares estimation for any multiple regression

\[
y_i = \beta_0 + \sum_{k=1}^{p} x_{ik}\beta_k + \epsilon_i, \quad i = 1, \ldots, n,
\]

is equivalent to

\[
y_i = \bar{y} + \sum_{k=1}^{p} (x_{ik} - \bar{x}_k)\beta_k + \epsilon_i,
\]
i.e. we must have

\[ \hat{\beta}_0 = \bar{y} - \sum_{k=1}^{p} \bar{x}_k \hat{\beta}_k. \]

Assume all predictors \( X_k \) have been centered, i.e. after projection into the constant vector 1, which corresponds to the intercept. When all the variables are uncorrelated, \( X'X \) is a diagonal matrix. The regression coefficients do not depend on each other

\[ b = (X'X)^{-1} X'Y = \begin{bmatrix} \bar{Y} \\ X'_1Y / \|X_1\|^2 \\ \vdots \\ X'_pY / \|X_p\|^2 \end{bmatrix}, \]
where

\[ X_k'Y = \sum_{i=1}^{n} x_{ik}y_i, \quad \|X_k\|^2 = \sum_{i=1}^{n} x_{ik}^2, \]

and hence we can decompose the \(SSR\) into individual predictors

\[
SSR = \|X b - \bar{Y}\|^2 = b'_{-1} X'_{-1} X_{-1} b_{-1} = \sum_{k=1}^{p} \|X_kb_k\|^2 = \sum_{k=1}^{p} \sum_{i=1}^{n} b_k^2 x_{ik}^2,
\]

where \(\bar{Y} = b_0 1\), \(b_{-1}\) does not contain the first regression coefficient, i.e. the intercept, and \(X_{-1}\) does not contain the first column, which corresponds to the intercept.

So we have

\[
SSR(X_S|X_F) = SSR(X_S), \tag{13}
\]

where \(S\) and \(F\) are two disjoint sets of indices.
Correlated Predictors

If $X_k = X_l$, for any constant $c_0$, $(\hat{\beta}_k - c_0, \hat{\beta}_l + c_0)$ will also be the least squares solutions, but the error sum of squares are still the same. Essentially we can only estimate $\beta_k + \beta_l$.

So the perfect relation between $X_k$ and $X_l$ does not prevent us from obtaining a good fit to the data. But it is meaningless to interpret the regression coefficients in this case since they have infinite solutions.

In practice due to random noise in the data, perfect relation seldom occurs. But generally big correlations among predictors will cause big variability for the regression coefficient estimations.

We can use regression to decompose $X_k$ into the sum of some linear
functions of \{X_j, j \neq k\} and residuals vector, which is orthogonal to \{X_j, j \neq k\} and \mathbf{1}:

\[ X_k = \theta_0 + \sum_{j \neq k} X_j \theta_j + e_k. \]

Then we can rewrite the multiple regression as following

\[ Y = \beta_0 + \theta_0 \beta_k + \sum_{j \neq k} X_j (\beta_j + \beta_k \theta_j) + e_k \beta_k + \epsilon. \]

Because \( e_k \perp X_j, j \neq k \) and \( e_k \perp \mathbf{1} \), we have

\[ \hat{\beta}_k = \frac{\sum_{i=1}^n e_{ik}y_i}{\sum_{i=1}^n e_{ik}^2}, \]
and

$$Var(\hat{\beta}_k) = \frac{\sigma^2}{\sum_{i=1}^{n} e_{ik}^2}.$$  

If $X_k$ has big correlation with anyone of the $X_j, j \neq k$, the residuals will be very small. Formally we have the following inequality

$$\sum_{i=1}^{n} e_{ik}^2 \leq (1 - r_{kl}^2) \sum_{i=1}^{n} x_{ik}^2, \quad \forall l \neq k.$$  

Some effects of big correlation on the regressions (read the textbook on those numerical examples)

- It does not prevent good fit of the data. So the MSE is not affected too much, neither are the inference about mean response and new observation.
Our previous analysis shows that the regression coefficient estimations tend to have big variance, which will make the inference instable.

The extra sum of squares is not a good measure of variable contribution when big correlation exists.

Regression parameter estimation depends on other variables in the model. The simultaneous testing of regression coefficients are also dependent on each other.
Polynomial Regression

Polynomial regression is usually used to approximate the underlying unknown curvilinear relations. Polynomial regression is global model and extrapolation is usually not a good idea.

For one predictor we can generally consider \( p \)-order model

\[
Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_i^k + \epsilon_i,
\]

where we usually use centered predictor, i.e. \( x_i = X_i - \bar{X} \), to reduce the correlation among different order terms. \( \beta_1 \) is often called the linear effect coefficient, \( \beta_2 \) the quadratic effect coefficient etc.
When there are more than one predictor in the model, we can consider their interactions, e.g.

\[ Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \epsilon_i, \]

where \( \beta_{12} \) is often called the \textit{interaction effect coefficient}. 
Interaction Regression Model

Consider previous interaction model

\[ E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2, \]

we have

\[ \frac{\partial E(Y)}{\partial X_1} = \beta_1 + \beta_{12} X_2, \quad \frac{\partial E(Y)}{\partial X_2} = \beta_2 + \beta_{12} X_1, \]

so the mean response change per unit change in \( X_1 \) (\( X_2 \)) now depends on other variable \( X_2 \) (\( X_1 \)).
Constrained Regression

Sometimes we have some prior information about some regression coefficients. We can incorporate this information into our model to do constrained regression.

Consider regression example

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

the information of $\beta_1 = 1$ is relatively easy to incorporate into the model, we just need to do the following simple regression

$$\mathbb{E}(Y - X_1) = \beta_0 + \beta_2 X_2.$$
But for other types of information it is hard to do the constrained regression, e.g. if we constrain $\beta_1 > \beta_2$ for previous model. We often have to resort to numerical optimization technique to solve the solutions.