Additional Topics in MLR

1 QR Decomposition: How the LS estimates $\hat{\beta}$ are solved in R

Denote the QR decomposition (also called QR factorization) of $\mathbf{X}$ as

$$\mathbf{X}_{n\times p} = \mathbf{Q}_{n\times p}\mathbf{R}_{p\times p}$$

where $\mathbf{Q}$ is an orthogonal matrix (i.e. $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_{p\times p}$) and $\mathbf{R}$ is an upper triangular matrix, i.e. all the entries in $\mathbf{R}$ below the diagonal are equal to 0. Then,

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \\
(\mathbf{X}^T\mathbf{X})^{-1} &= (\mathbf{R}^T\mathbf{R})^{-1} = \mathbf{R}^{-1}(\mathbf{R}^T)^{-1} \\
\hat{\beta} &= \mathbf{R}^{-1}\mathbf{Q}\mathbf{y} \\
\mathbf{R}\hat{\beta} &= \mathbf{Q}\mathbf{y}
\end{aligned}$$

The last equation is solved easily via backsolving since $\mathbf{R}$ is an upper triangular matrix.

Gram-Schmidt Algorithm

One method for computing the QR decomposition is the Gram-Schmidt algorithm. It works as follows: Take

$$\mathbf{A}_{n\times p} = [a_1|a_2|\dots|a_p],$$

where a_j denotes the jth column of $\mathbf{A}$. Then, define a sequence of e_i's and $\mathbf{q}_i's$ recursively:

(1) $e_1 = a_1$, $\mathbf{q}_1 = \frac{e_1}{||e_1||}$

(2) $e_2 = a_2 - (a_2^T\mathbf{q}_1)\mathbf{q}_1$, $\mathbf{q}_2 = \frac{e_2}{||e_2||}$

(3) ...

(4) $e_{k+1} = a_{k+1} = \sum_{j=1}^{k}(a_j^T\mathbf{q}_j)\mathbf{q}_j$

The resulting QR decomposition is

$$\mathbf{A}_{n\times p} = [a_1|a_2|\dots|a_p] = \mathbf{A}_{n\times p} = [\mathbf{q}_1|\mathbf{q}_2|\dots|\mathbf{q}_p]\mathbf{R} = \mathbf{QR}$$

It is easy to check that $\mathbf{R}$ is indeed an upper triangular matrix.

2 Partial Regression Coefficients

Consider a multiple linear regression model with four predictors and an intercept

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

The LS estimate $\hat{\beta}_k$ describes the partial correlation between Y and X_k adjusted for the other predictors. Mathematically, the LS estimate $\hat{\beta}_k$ is what we could get if we

(1) first regress Y onto all other predictors except X_k, denote the corresponding residuals as a new variable Y^*

(2) regress X_k onto all other predictors except X_k, denote the corresponding residuals as a new variable X_k^*

(3) then fit a simple linear regression model with Y^* as the response and X_k^* as the predictor.

As an example, let us take the `savings`[1] data set from the *faraway* library we have:

```
>fullmodel=lm(sr~pop15+pop75+dpi+ddpi, data=savings)
> round(summary(fullmodel)$coef, dig=3)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   28.566      7.355   3.884    0.000
pop15         -0.461      0.145  -3.189    0.003
pop75         -1.691      1.084  -1.561    0.126
dpi            0.000      0.001  -0.362    0.719
ddpi           0.410      0.196   2.088    0.042

>new.y=lm(sr~pop15+pop75+dpi, data=savings)$res
>new.ddpi = lm(ddpi~pop15+pop75+dpi, data=savings)$res
>parmodel=lm(new.y ~ new.ddpi)
>round(summary(parmodel)$coef, dig=3)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.00      0.521   0.000    1.000
new.ddpi        0.41      0.190   2.157    0.036
```

Note that although the estimated coefficients agree, the standard error and p-value for the corresponding t-test are different. This is because the sample size is miscounted in the 2nd regression: the sample size should not be be n, but n-4 since both the regressor Y^* and the predictor X^* are in an n-4 subspace which is orthogonal to the intercept and the other 3 predictors.

```
>sd=summary(parmodel)$coef[2,2]*sqrt(48/45) # correct std. error
>sd
[1] 0.1961971
>2*(1-pt(abs(summary(parmodel)$coef[2,1]/sd), 45)) # correct p-value
[1] 0.04247114
```

Since the t-test for β_k is testing the effect of X_k adjusted for the other predictors, it is not surprising to see the equivalence of the t-test and the F -test of comparing the full model to the model including all predictors except X_k.

3 Sequential analysis of variance (ANOVA)

When adding a new predictor to a regression model, we can evaluate its relevance/EFFECT by the improved RSS. When there are multiple predictors, we can carry out this comparison in a sequential

[1] You can find more information regarding this data set with `help(savings)` after loading the *faraway* library.

way: FIRST compare the RSS from the model with X_1 to the RSS from the null model (with only the intercept), then compare the RSS from the model with both X_1 and X_2 to the RSS from the model with X_1 only, and so on. That IS what the R command anova would return you.

```
>rss = sum((sr-mean(sr))^2)
>model1 = lm(sr~pop15)
>rss = c(sum(model1$res^2), rss)
>model2 = lm(sr~pop15+pop75)
>rss = c(sum(model2$res^2), rss)
>model3 = lm(sr~pop15+pop75+dpi)
>rss = c(sum(model3$res^2), rss)
>model4 = lm(sr~pop15+pop75+dpi+ddpi)
>rss = c(sum(model4$res^2), rss)
>rss
[1] 650.7130 713.7670 726.1680 779.5107 983.6283
>round(diff(rss), dig=2)
[1] 63.05 12.40 53.34 204.12
>anova(fullmodel)
Analysis of Variance Table

Response: sr
          Df Sum Sq Mean Sq F value    Pr(>F)
pop15      1 204.12 204.118 14.1157 0.0004922 ***
pop75      1  53.34  53.343  3.6889 0.0611255 .
dpi        1  12.40  12.401  0.8576 0.3593551
ddpi       1  63.05  63.054  4.3605 0.0424711 *
Residuals 45 650.71  14.460

> anova(lm(sr~ddpi+pop15+pop75+dpi))

Analysis of Variance Table

Response: sr
          Df Sum Sq Mean Sq F value    Pr(>F)
ddpi       1  91.37  91.374  6.3190 0.0155920 *
pop15      1 191.70 191.702 13.2571 0.0006984 ***
pop75      1  47.95  47.946  3.3157 0.0752748 .
dpi        1   1.89   1.893  0.1309 0.7191732
Residuals 45 650.71  14.460
```

Of course, order matters: the importance of ddpi as the 1st variable entering the model wouldn't be expected to be the same as the last variable entering the model.

We have already seen that evaluating the effect of a single predictor is a difficult problem, since its effect depends on what else are included in the model. Later in this semester we will learn how to select an optimal subset of variables.

4 Error in Predictors

Let $\mathbf{X}$ be the observed design matrix of dimension $n \times p$ where the 1st column contains only 1's and the remaining $(p-1)$ columns correspond to the $(p-1)$ non-intercept covariates (predictors). In many cases it is quite possible that there are substantial measurement errors involved. Let $\tilde{X}$ denote the "true" value of the predictors, and consider the model

$$\mathbf{X} = \tilde{\mathbf{X}} + \mathbf{D}$$

where $\mathbf{D}$ is a matrix of errors of the same dimension as $\mathbf{X}$. When an intercept is in the model, the first column of $\mathbf{D}$ would contain only 0's. The remaining elements of $\mathbf{D}$ could represent rounding errors or measurement errors.

Let d_i^T denote the row of $\mathbf{D}$ corresponding to the$i$-th case (so d_i is a $p \times 1$ vector). We assume that d_i and d_j are statistically independent for $i \neq j$. In addition, we assume that

$$\mathbf{E}(d_i) = 0,\ Cov(d_i) = \mathbf{S} = diag(s_i^2)_{i=1}^p$$

where $s_1^2 = 0$ (i.e. no measurement error for the intercept).

Suppose the problem of interest is to estimate β in the model

$$y = \tilde{X}\beta + e$$

We would like to estimate β with $(\tilde{\mathbf{X}}^T(\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T y$. However, we observe $\mathbf{X}$ instead of $\tilde{\mathbf{X}}$. Instead, we have

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

When the predictors are measured with error, the LS estimator $\hat{\beta}$ obtained in is no longer unbiased. The bias is given by

$$\mathbb{E}(\hat{\beta}) - \beta = -(n-p-1)(\tilde{\mathbf{X}}^T(\tilde{\mathbf{X}})^{-1}\mathbf{S}\beta$$

With a given data set, this can be approximated by

$$\mathbb{E}(\hat{\beta}) - \beta \approx -(n-p-1)(\tilde{\mathbf{X}}^T(\tilde{\mathbf{X}})^{-1}\mathbf{S}\hat{\beta}$$

In the special case of estimating the slope in simple linear regression, the approximation is

$$\mathbb{E}(\hat{\beta}_1) \approx \beta_1 \left[1 - \frac{s_1^2}{\sum_{i=1}^n (x_i - \bar{x})^2/(n-2)}\right]$$