36-401 Variable Selection (part 2)

The following quotes from Weisberg's Applied Linear Regression are better than anything in RwG on this subject.

Deletion of predictors from a model can improve a model and reduce apparent collinearity. Consider the choice between the models

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \quad \epsilon \sim N(0, \sigma^2) \tag{eq1}$$

with p=2 predictors, and the subset model with p=1 predictor

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \tag{eq2}$$

obtained from (eq1) by deleting variable X_2. Suppose the estimation of β_1 is of interest. We assume that model (eq1) is the correct description of the dependence of Y on the predictors, and that errors are independent with constant variance. For simplicity we also assume that $\sum(X_1 - \bar{X}_1)^2$=$\sum(X_2 - \bar{X}_2)^2$=1. For model (eq1), the usual least squares estimate of β_1 is unbiased, and ...

$$\text{Var}\left[\hat{\beta}_1 \mid \text{full model (eq1)}\right] = \frac{\sigma^2}{1 - r_{12}^2}.$$

For model (eq2), more care is required, since β_1 now has a different meaning: it is the effect of X_1 on Y ignoring X_2, while in (eq1) it is the effect of X_1 on Y adjusting for X_2. Consequently, if these two are not the same, then the estimate of β_1 from (eq2) will be biased; one can show that

$$\text{E}(\hat{\beta}_1 \mid \text{subset model}) = \beta_1 + r_{12}\beta_2.$$

...the variance of $\hat{\beta}_1$ from the subset model can be shown to be

$$\text{Var}(\hat{\beta}_1 \mid \text{subset model}) = \sigma^2$$

which does not depend on r_{12}. To compare the estimation of β_1 from the full and subset models, we must compute the mean squared error of $\hat{\beta}_1$.... Thus when collinearity is observed in the data, more precise estimation of β_1 can generally be obtained from the subset model than the full model, unless the coefficient for the the deleted variable is very large.

[. . .]

The single most important tool in selecting a subset of variables for use in a model is the analyst's knowledge of the substantive area under study and of each of the variables, including expected sign and magnitude of the coefficient.

[. . .]

Another important method of reducing the number of predictors is to define new variables that are combinations of the old variables. For example, in some studies, the two variables height and weight, which are often highly correlated, may be replaced by a combination of them to give a single height-weight index, or two IQ tests as predictors can be replaced by their average.

In many problems, a point is reached at which it is necessary to use the data to find subsets of the predictors. Two basic methods for this purpose have evolved. The first method, generally called *stepwise* regression, uses a convenient computational algorithm to limit possible models to a relatively small number. It is very heavily used in practice, although it does not correspond to any specific criterion for choosing a model. The second method uses a criterion statistic for all possible subsets of predictors. This has become practical with the development of several fast algorithms for computing all possible regressions.

t-tests and CI's for Individual Coefficients

estSE_{b_k} is in summary(lm()) in Splus. This is just the appropriate diagonal element of $\mathbf{S} = s_e^2\,(\mathbf{X}'\mathbf{X})^{-1}$.

To test $H_0 : \beta_k = 0$, calculate $t = \frac{b_k - \beta_{khyp}}{\text{estSE}_{b_k}}$ and compare to a t-distribution. The p-value (for a two sided alternative hypothesis) in Splus is 2*(1-pt(tval,df)).

The confidence limit for β_k is

$$b_k \pm (t)(\text{estSE}_{b_k}),$$

where t is the appropriate quantile of the t-distribution. For type 1 error of α, e.g. 0.05, the t-value can be obtained in Splus as qt(1-α/2, df).

t-tests and CI's for $\hat{Y}^*$

$$\text{E}(\hat{Y}^*) = \text{E}(b_0 + b_1 X_1^* + b_2 X_2^*) = \beta_0 + \beta_1 X_1^* + \beta_2 X_2^*$$

$$\begin{aligned}
\text{Var}(b_0 + b_1 X_1^* + b_2 X_2^*) &= \text{Var}(b_0) + 2X_1^*\text{Cov}(b_0, b_1) + 2X_2^*\text{Cov}(b_0, b_2) \\
&+ (X_1^*)^2\text{Var}(b_1) + (X_2^*)^2\text{Var}(b_2) + 2X_1^*X_2^*\text{Cov}(b_1, b_2) \\
\cdots &= \mathbf{x}^{*\prime}\,\mathbf{S}\,\mathbf{x}^*
\end{aligned}$$

where $\mathbf{x}^{*\prime} = (1\ x_1^*\ x_2^*)$.

Add s_e^2 for prediction variance. (Then) take square root for standard error. Confidence intervals as usual, are $\hat{Y}^* \pm t \cdot \text{SE}$.

F-tests for Sets of Coefficients

These are tests for nested hypotheses, i.e. we are comparing two models, one of which is a special case of the other. The usual special case is that one or more (H) of the coefficients are set to zero, so this is a comparison of a larger and a smaller model. These F-tests are likelihood ratio tests.

Call the larger model "full" and the smaller one "reduced". We want to test the null hypothesis that the H dropped coefficient(s) are (all) zero.

The test statistic is

$$F_{n-K}^{H} = \frac{\left(\mathrm{RSS}_{\mathrm{red}} - \mathrm{RSS}_{\mathrm{full}}\right)/H}{\left(\mathrm{RSS}_{\mathrm{full}}\right)/(n-K)}$$

where $\mathrm{RSS}_{\mathrm{red}}$ is the sum of the squared residuals from the reduced model and $\mathrm{RSS}_{\mathrm{full}}$ is the sum of the squared residuals from the full model. F_{n-K}^{H} is compared to an F-distribution with H and n-K df.

We can get the F and p values in S-plus with anova():

```
> lm.red _ lm(LogWage ~  Educ + South + Sex + Exper,data=wages)
> lm.full _ lm(LogWage ~  Educ + South + Sex + Exper + Union + Sector,
+ data=wages)
> anova(lm.red,lm.full)
Analysis of Variance Table

Response: LogWage

                                        Terms Resid. Df      RSS          Test
1                   Educ + South + Sex + Exper       529 106.6999
2 Educ + South + Sex + Exper + Union + Sector       527 102.6309 +Union+Sector

  Df Sum of Sq  F Value        Pr(F)
1
2  2  4.069059 10.44712 3.552278e-05
```

Think of the F test as a check whether the drop in residual sum of squares due to adding some covariates is large enough to be worth while ($p < \alpha$) or so small as to suggest that the covariates are useless ($p > \alpha$).

This can be modified to test general hypotheses, i.e.

$H_0 : \mathbf{A}\boldsymbol{\beta}_Z = \mathbf{C}$ where $\boldsymbol{\beta}_Z$ represents the betas for the covariates under special consideration.

Two other ways to get the same F test:

- By looking at RSS and RMS values in `summary()` applied to each model.

```
> summary(lm.full)
[...]

Residual standard error: 0.4413 on 527 degrees of freedom
Multiple R-Squared: 0.3086
F-statistic: 39.21 on 6 and 527 degrees of freedom, the p-value is 0

[...]
> 0.4413^2 * 527
[1] 102.631

> summary(lm.red)
[...]

Residual standard error: 0.4491 on 529 degrees of freedom
Multiple R-Squared: 0.2812
F-statistic: 51.74 on 4 and 529 degrees of freedom, the p-value is 0

[...]
>  .4491^2 * 529
[1] 106.6944

> ( (106.6999-102.6309)/2 ) / ( 102.6309/527 )
[1] 10.44697
> 1-pf(10.44697,2,527)
[1] 3.552795e-05
```

- By arranging to have the terms by which the models differ appear last in the `anova()` output.

```
> anova(lm.full)

Terms added sequentially (first to last)
           Df Sum of Sq  Mean Sq  F Value      Pr(F)
     Educ   1   21.4807 21.48068 110.3013 0.00000000
    South   1    2.1170  2.11704  10.8708 0.00104277
      Sex   1    7.3098  7.30979  37.5351 0.00000000
    Exper   1   10.8394 10.83939  55.6593 0.00000000
    Union   1    2.9445  2.94455  15.1200 0.00011378
   Sector   1    1.1245  1.12451   5.7743 0.01660726
Residuals 527  102.6309  0.19475
> ( (2.9445+1.1245)/2 ) / ( 102.6309 / 527 )
[1] 10.44697
> 1-pf(10.44697,2,527)
[1] 3.552795e-05
```

Breakout Session – Variable Selection II

`grass.dat` is from Linthurst, R.A. (1979) "Aeration, nitrogen, pH, and salinity as factors affecting Spartina alerniflora growth and dieback," Ph.D. thesis. North Carolina State University. The variables are:

bio	biomass i(gm per square meter)
sal	salinity (parts per thousand)
K	potassium (parts per million)
Na	sodium (parts per million)
Zn	zink (parts per million)

```
> grass _ read.table("grass.dat",header=T)
> summary(lm(bio~sal+pH+K+Na+Zn,grass))

Coefficients:    Value Std. Error    t value   Pr(>|t|)
(Intercept)  1252.5147  1234.7474     1.0144     0.3166
        sal   -30.2860    24.0309    -1.2603     0.2151
         pH   305.4631    87.8827     3.4758     0.0013
          K    -0.2849     0.3485    -0.8176     0.4185
         Na    -0.0087     0.0159    -0.5443     0.5893
         Zn   -20.6801    15.0535    -1.3738     0.1774

Residual standard error: 398.3 on 39 degrees of freedom
Multiple R-Squared: 0.6773
F-statistic: 16.37 on 5 and 39 degrees of freedom, the p-value is 1.082e-08

Correlation of Coefficients:
    (Intercept)     sal      pH       K      Na
sal -0.9180
 pH -0.8406      0.6045
  K -0.3123      0.3047  0.1213
 Na  0.3768     -0.4324 -0.2442 -0.8105
 Zn -0.8809      0.7114  0.8350  0.1909 -0.3387

> summary(lm(bio~sal+pH+K,grass))

Coefficients:   Value Std. Error   t value  Pr(>|t|)
(Intercept) -131.3064  582.5284    -0.2254    0.8228
        sal  -12.0501   16.3693    -0.7361    0.4658
         pH  410.1419   48.8273     8.3998    0.0000
          K   -0.4899    0.2043    -2.3977    0.0211

Residual standard error: 403.3 on 41 degrees of freedom
Multiple R-Squared: 0.6522
F-statistic: 25.62 on 3 and 41 degrees of freedom, the p-value is 1.685e-09

Correlation of Coefficients:
    (Intercept)     sal      pH
sal -0.8756
 pH -0.4242      0.0510
  K -0.2895      0.0195 -0.0176
```

1. Why did the coefficients change between models?

2. Is potassium (K) significant?

3. Test the hypothesis $H_0 : \beta_{Na} = \beta_{Zn} = 0$, i.e. show how to get the F value from the anova().

4. What is the 95% CI for predicted biomass for a future observation with sal=30, pH=5, K=1000?
   $t_{(0.05,41)}$=2.02
   predict(grasslm, data.frame(sal=30, pH=5, K=1000), se.fit=T) gives:
   $fit: 1067.99
   $se.fit: 75.34647
   $df: 41

   If you have time, show how to find se.fit from the data in the summary(lm()).