Regression Review

Statistics 149

Spring 2006

Matrix Approach to Regression

Linear Model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \ i = 1, \ldots, n$$

$$\begin{aligned} Y_i | X_{i1}, \ldots, X_{ip} &\overset{ind}{\sim} N(\mu_i, \sigma^2) \\ \mu(Y_i | X_{i1}, \ldots, X_{ip}) &= \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} = \mu_i \\ \mathrm{Var}(Y_i | X_{i1}, \ldots, X_{ip}) &= \sigma^2 \end{aligned}$$

For what follows, the inclusion of the intercept will be assumed, though it need not be.

This model can be written in matrix notation as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

- Responses $\mathbf{Y}$: $n \times 1$ (rows $times$ cols)

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

- Predictors $\mathbf{X}$: $n \times (p+1)$

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}$$

- Regression parameters $\boldsymbol{\beta}$: $(p+1) \times 1$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

- Errors $\boldsymbol{\epsilon}$: $n \times 1$

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The corresponding distributional assumptions in the matrix formulation are

$$\begin{aligned} \boldsymbol{\epsilon} &\sim N_n(\mathbf{0}, \sigma^2 I_n) \\ \mathbf{Y}|\mathbf{X} &\sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n) \\ \mu(\mathbf{Y}|\mathbf{X}) &= \mathbf{X}\boldsymbol{\beta} \\ \mathrm{Var}(\mathbf{Y}|\mathbf{X}) &= \sigma^2 I_n \end{aligned}$$

where $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the n-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$ and I_n is the $n \times n$ identity matrix (Note I'll often not indicate the dimension of the identity matrix).

In what follows, it will be assumed that $\mathrm{rank}(\mathbf{X}) = p + 1$, which will lead to a unique solution of least squares criterion

$$\begin{aligned} SS(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2 \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

with respect to $\boldsymbol{\beta}$. One way of thinking of this, is that no predictor is a linear combination of the rest.

The least squares solutions for $\boldsymbol{\beta}$ satisfy

$$\begin{aligned} \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}^T\mathbf{Y} \qquad \text{(Normal Equations)} \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \end{aligned}$$

The vector of fitted values satisfy

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= H\mathbf{Y}\end{aligned}$$

Geometrically, $\hat{\mathbf{Y}}$ can be thought of the projection of $\mathbf{Y}$ onto the subspace generated by the columns of $\mathbf{X}$.

The matrix H is known as the hat matrix (a $n \times n$ matrix) and is important for many regression calculations and diagnostics (will come back to later).

The vector of residuals satisfy

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

Theorem. *Assume that* $\mathrm{Var}(\mathbf{Y}) = \Sigma$ *is a* $k \times k$ *matrix. Then if* A *is a* $l \times k$ *matrix, then* $\mathbf{Z} = A\mathbf{Y}$ *has variance matrix*

$$\mathrm{Var}(\mathbf{Z}) = A\Sigma A^T$$

Note that if A *is a vector (*$l = 1$*), when you work out the matrix multiplication, the result is equivalent to*

$$\mathrm{Var}(Z) = \sum_{i=1}^{k} a_i^2 \mathrm{Var}(Y_i) + \sum_{i<j} 2a_i a_j \mathrm{Cov}(Y_i, Y_j)$$

This theorem is the multivariate analogue of $\mathrm{Var}(aY) = a^2\mathrm{Var}(Y)$.

Some important variance results based on this result are

1. $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

2. $\mathrm{Var}(\hat{\mathbf{Y}}) = \sigma^2 H$

3. $\mathrm{Var}(\mathbf{e}) = \sigma^2(I - H)$

4. $\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \mathbf{e}) = \mathbf{0}_{(p+1)\times n}$

5. $\mathrm{Cov}(\hat{\mathbf{Y}}, \mathbf{e}) = \mathbf{0}_{n\times n}$

assuming that $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 I$ (constant variance and uncorrelated).

The first of these is justified by

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

The others are justified by the facts

- $H = H^T$
- Since H is a projection matrix, $H^2 = HH = H$.

Another implication of this second fact is that all the linear information in $\mathbf{Y}$ described by $\mathbf{X}$ is given by the linear regression.

Lets think about situation where we regress $\mathbf{Y}$ on $\mathbf{X}$, which gives residuals $\mathbf{e}$. Now lets regress $\mathbf{e}$ on $\mathbf{X}$, i.e fit the model

$$\mathbf{e} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}^*$$

and see what the fits and residuals for this second regression are.

- Estimate of $\boldsymbol{\beta}^*$:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^* &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(I - H)\mathbf{Y} \\
&= ((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y} = \mathbf{0}
\end{aligned}$$

- Fits of residuals:

$$\begin{aligned}\hat{\mathbf{e}} &= H\mathbf{e} \\ &= H(I-H)\mathbf{Y} \\ &= H\mathbf{Y} - H^2\mathbf{Y} \\ &= H\mathbf{Y} - H\mathbf{Y} = \mathbf{0}\end{aligned}$$

- Residuals of residuals:

$$\begin{aligned}\mathbf{r} = \mathbf{e} - \hat{\mathbf{e}} &= (I-H)\mathbf{e} \\ &= (I-H)(I-H)\mathbf{Y} \\ &= (I-H-H+H^2)\mathbf{Y} \\ &= (I-H)\mathbf{Y} = \mathbf{e}\end{aligned}$$

Now lets look at the Puffin example from last time to numerically display the results implied by these matrix calculations.

```
> puffin.lm <- lm(nesting ~ grass + soil + angle + distance,
    data=puffin)

> puffin.res <- resid(puffin.lm)

> summary(puffin.lm)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0166 -2.1088  0.2293  1.2505  6.9881

Residual standard error: 2.647 on 33 degrees of freedom
Multiple R-Squared: 0.8792,     Adjusted R-squared: 0.8645
F-statistic: 60.03 on 4 and 33 DF,  p-value: 1.113e-14
```

```
> resid.lm <- lm(puffin.res ~ grass + soil + angle + distance,
    data=puffin)
> summary(resid.lm)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0166 -2.1088  0.2293  1.2505  6.9881

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept)  1.700e-16  3.185e+00  5.34e-17        1
grass        3.997e-19  1.946e-02  2.05e-17        1
soil         6.538e-18  7.724e-02  8.47e-17        1
angle       -1.646e-17  7.780e-02 -2.12e-16        1
distance    -4.346e-18  5.747e-02 -7.56e-17        1

Residual standard error: 2.647 on 33 degrees of freedom
Multiple R-Squared:  4.699e-33, Adjusted R-squared: -0.1212
F-statistic:  3.877e-32 on 4 and 33 DF,  p-value: 1
```

```
> max(abs(fitted(resid.lm)))
[1] 4.440892e-16

> max(abs(puffin.res - resid(resid.lm)))
[1] 4.440892e-16
```

So the estimated βs and fitted residuals from this second regression are all 0, as is the difference between the residuals from the two regressions (up to the numerical precision of **R**).

So an implication of these general results, is that if you see pattern in the residuals, say for example some curvature in the plot of residuals vs one of the predictors, you will need to fit a different model.

This could be done by transforming Y, adding new predictors, or transforming current predictors (add in X^2 for example).

The usual estimate of σ^2 can be written as

$$\hat{\sigma}^2 = \frac{\mathbf{e}^T\mathbf{e}}{n-p-1} = \frac{\mathbf{Y}^T(I-H)\mathbf{Y}}{n-p-1} = MSE$$

Note that all of the results seen so far only depend on the moment assumptions

$$\mu(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$
$$\mathrm{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2 I_n$$

and not on normality. If we are willing to assume that

$$\mathbf{Y}|\mathbf{X} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n)$$

then due to

Theorem. *Let* $\mathbf{Y} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and let* A *be an* $k \times l$ *matrix. Then* $Z \sim N_l(A\boldsymbol{\mu}, A\boldsymbol{\Sigma}A^T)$.

the following additional results hold

1. $\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$

2. $X_0\hat{\boldsymbol{\beta}} \sim N_1(X_0\boldsymbol{\beta}, \sigma^2 X_0(\mathbf{X}^T\mathbf{X})^{-1}X_0^T)$. $X_0\hat{\boldsymbol{\beta}}$ is the estimated mean response for predictor vector X_0.

3. $\hat{\mathbf{Y}} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 H)$

4. $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2(I - H))$

5. $SSE = \mathbf{e}^T\mathbf{e} = \mathbf{Y}^T(I - H)\mathbf{Y} \sim \sigma^2\chi^2_{n-p-1}$

6. $\hat{\boldsymbol{\beta}}$ is independent of $\hat{\sigma}^2$ since $\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \mathbf{e}) = \mathbf{0}$

The 1st, 2nd, and 5th results are needed to justify the t procedures used on $\hat{\beta}$ and for prediction intervals and confidence intervals for mean response, e.g.

$$\frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)} \sim t_{n-p-1}$$

In addition, the components of the usual sums of squares decomposition

$$SST = SSR + SSE$$

can be easily calculated by

$$SSR = Y^T \left[H - \frac{1}{n} J \right] Y$$
$$SSE = Y^T (I - H) Y$$
$$SST = Y^T \left[I - \frac{1}{n} J \right] Y$$

where J is a $n \times n$ matrix of all 1's.

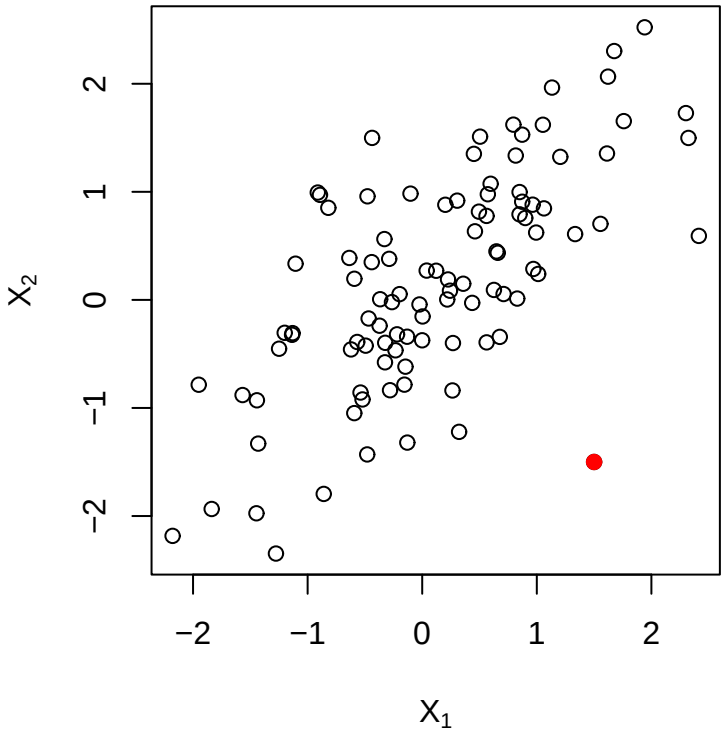
Regression Diagnostics

As mentioned earlier, components of the Hat matrix H can be used for many diagnostic purposes.

The components of the vector $h = \mathrm{diag}(H)$ are known as the leverages. Note that in the case when there is only a single predictor, these values correspond to the well known formula

$$h_i = \frac{1}{n-1}\left[\frac{X_i - \bar{X}}{s_x}\right]^2 + \frac{1}{n} = \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2} + \frac{1}{n}$$

The leverages measure how far away the predictors are away from the average of the predictors, accounting for the correlation amongst the predictors, on a relative scale. An observation could have a high leverage (such as the red point), even though it is not extreme for each variable individually.



It can be shown that

$$\frac{1}{n} \leq h_i \leq 1$$

and

$$\sum_{i=1}^{n} h_i = p + 1$$

(assuming that there is an intercept in the model).

By examining the structure of the matrix multiplication $\hat{\mathbf{Y}} = H\mathbf{Y}$, it is evident that

$$\hat{Y}_j = h_{jj}Y_j + \sum_{i \neq j} h_{ji}Y_i$$

(where h_{ji} is the element from row j and column i of H) so the leverages give some information about how much information each observation has on its own fit (as $h_{jj} = h_j$). These influence ideas can be expanded on.

For example, based on the earlier variance results,

$$\mathrm{Var}(\hat{Y}_i) = \sigma^2 h_i$$
$$\mathrm{Var}(e_i) = \sigma^2(1 - h_i)$$

So values with high leverages will tend to have smaller residuals.

This is the reason that studentized residuals

$$studres_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}$$

are usually better to check for possible outliers.

Another alternative for checking for outliers are deleted residuals, which are given by

$$d_i = Y_i - \hat{Y}_{i(i)}$$

where $\hat{Y}_{i(i)}$ is the fit of observation i when it is not included in the fitting of the model (use the other $n-1$ observations to estimate $\boldsymbol{\beta}$.)

It ends up you don't need to rerun the regression as

$$d_i = \frac{e_i}{1-h_i}$$

where e_i and h_i are taken from the full regression.

To check for outliers, the studentized deleted residual

$$t_i = \frac{d_i}{SE(d_i)} = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}$$

can be compared to critical values a t_{n-p-1} distribution. $\hat{\sigma}^2_{(i)}$ is the estimate of σ^2 when observation i is not used for fitting. It can be calculated from the main regression by the relationship

$$(n-p-1)\hat{\sigma}^2 = (n-p-2)\hat{\sigma}^2_{(i)} + \frac{e_i^2}{1-h_i}$$

In addition, h can be used to search for potentially influential observations. For example, values of $h_i > \frac{2(p+1)}{n}$ are often considered as outliers in their X values.

In addition, Cook's Distance, one common measure of influence depends on h

$$D_i = \sum_{i=j}^{n} \frac{(\hat{Y}_{j(i)} - \hat{Y}_j)^2}{(p+1)\hat{\sigma}^2}$$

$$= \frac{e_i^2}{(p+1)\hat{\sigma}^2} \frac{h_i}{(1-h_i)^2}$$

where $\hat{Y}_{j(i)}$ is the jth fitted value from a fit that excludes observation i.

Cook's distance measures the overall influence an observation has on the overall fit.

The second version of the formula shows that this measure depends on the residual and the leverage.

An observation could have a large D_i value if

- Large residual and a moderate leverage
- Large leverage and moderate residual
- Large residual and leverage

D_i larger that 1 are often indicative of large influence. A D_i which is much larger than the rest, but less than 1, is also an indicator.

Another useful measure of influence is DFFITS, which just measures the influence that observation i has on its own fit. Its formula is given by

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_i}}$$

It can be more easily calculated by the formula

$$DFFITS_i = e_i \left[\frac{n-p-2}{SSE(1-h_i) - e_i^2}\right]^{1/2} \left(\frac{h_i}{1-h_i}\right)^{1/2} = t_i \left(\frac{h_i}{1-h_i}\right)^{1/2}$$

For small or moderate sized datasets, DFFITS exceeding 1 in magnitude are often considered large. For large datasets $2\sqrt{(p+1)/n}$ is the more usual cutoff.

Another measure of influence is DFBETAS, which measures the influence of each of the observations on the estimated βs. They are given by

$$DFBETAS_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\hat{\sigma}_{(i)}\sqrt{c_k}}; \quad k = 0, \ldots, p$$

where c_k is the kth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$.

This measures the influence of observation i on the estimate of β_k

Large values of DFBETAS are 1 for small or medium sized data sets and $2/\sqrt{n}$ for large ones.

Note that there is a tie between Cook's Distance and DFBETAS as

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{(p+1)\hat{\sigma}^2}$$

To get these various diagnostic measures in **R**, see `help(influence.measures)` for more details