

# **Diagnostics**

## Motivation

When working with a linear model with design matrix  $X$ , we may optimistically suppose that

$$EY \in \text{col}(X) \quad \text{and} \quad \text{var}(Y|X) = \sigma^2 I.$$

Point estimates and inferences depend on these assumptions approximately holding.

Inferences for small sample sizes may also depend on the distribution of  $Y - EY$  being multivariate Gaussian, but for moderate or large sample sizes this is not critical.

If point estimates and inferences are to be meaningful, we should check whether the data satisfy these assumptions.

## Residuals

The residuals can be expressed

$$R \equiv (I - P)Y$$

where  $P$  is the projection onto  $\text{col}(X)$ .

The residuals have two key mathematical properties regardless of the truth of the model specification:

- The residuals sum to zero, since  $(I - P)\mathbf{1} = 0$  and hence  $\mathbf{1}'R = \mathbf{1}'(I - P)Y = 0$ .
- The residuals and fitted values have zero sample covariance:

$$\begin{aligned}
 \widehat{\text{cov}}(R, \hat{Y}) &\propto (R - \bar{R})'\hat{Y} \\
 &= R'\hat{Y} \\
 &= Y'(I - P)PY \\
 &= 0.
 \end{aligned}$$

These properties hold as long as an intercept is included in the model (so  $P \cdot \mathbf{1} = \mathbf{1}$ , where  $\mathbf{1}$  is a vector of 1's).

If the basic linear model assumptions hold, these two properties have population counterparts:

- The expected value of each residual is zero:

$$\begin{aligned}ER &= (I - P)EY \\ &= 0.\end{aligned}$$

- The population covariance between any residual and any fitted value is zero:

$$\begin{aligned}\text{cov}(R, \hat{Y}) &= ER\hat{Y}' \\ &= (I - P)\text{cov}(Y)P \\ &= \sigma^2(I - P)P \\ &= 0.\end{aligned}$$

If the model is correctly specified, there is a simple formula for the variances and covariances of the residuals:

$$\begin{aligned}\text{cov}(R) &= (I - P) (EYY') (I - P) \\ &= (I - P) (X\beta\beta'X' + \sigma^2I) (I - P) \\ &= \sigma^2(I - P).\end{aligned}$$

If the model is correctly specified, the “standardized residuals”

$$\frac{Y_i - \hat{Y}_i}{\hat{\sigma}}$$

and the “Studentized residuals”

$$\frac{Y_i - \hat{Y}_i}{\hat{\sigma}(1 - P_{ii})^{1/2}}$$

approximately have mean zero and variance one.

## External standardization of residuals

Let  $\hat{\sigma}_{-i}^2$  be the estimate of  $\sigma^2$  obtained by fitting a regression model omitting the  $i^{\text{th}}$  case. It turns out that we can calculate this value without actually refitting the model:

$$\hat{\sigma}_{-i}^2 = \frac{(n - p - 1)\hat{\sigma}^2 - r_i/(1 - P_{ii})}{n - p - 2}$$

where  $r_i$  is the residual for the model fit to all data.

The “externally standardized” residuals are

$$\frac{Y_i - \hat{Y}_i}{\hat{\sigma}_{-i}},$$

The “externally Studentized” residuals are

$$\frac{Y_i - \hat{Y}_i}{\hat{\sigma}_{-i}(1 - P_{ii})^{1/2}}.$$

## Masking

Externally Studentized residuals solve the problem of a single large outlier inflating  $\hat{\sigma}$  and thereby “masking” itself.

If multiple large outliers may be present,  $IQR/1.35$  or  $MAD/0.65$  can be used to produce an alternative estimate of  $\sigma$  for standardizing residuals.

$IQR$  is the interquartile range of the residuals (the distance between the 25<sup>th</sup> and 75<sup>th</sup> percentiles).  $MAD$  is the median absolute value of the residuals.

## Use of residuals in model diagnostics

**Variability of the Studentized residuals:** comparing the Studentized residuals to a standard normal reference distribution, values greater than 2 should be rare (occurring less than one in twenty cases) and values greater than 3 should only occur in large data sets.

*Example:* Suppose that two covariates  $X_1$  and  $X_2$  are independent with mean zero and variance 1, and the data generating model is

$$Y = X_1 + X_2 + fX_1X_2 + \epsilon,$$

where  $\text{var}\epsilon = 1$ . Note that  $\text{var}(X_1X_2) = 1$ , and  $X_1X_2$  is uncorrelated with  $X_1$  and  $X_2$ . Therefore the interaction explains around  $f^2/(3 + f^2)$  of the variance in  $Y$ .

Using simulation with sample size 40, we can see that the range of the residuals grows with  $f$ , as does the number of residuals with magnitude greater than two (these are internally studentized residuals).

$f$	$\min(R)$	$\max(R)$	$\# > 2$
0.0	-2.2	2.2	1.7
0.5	-2.7	2.7	2.9
1.0	-2.7	2.8	3.0
1.5	-2.8	2.8	3.0
2.0	-2.8	2.8	3.0
2.5	-2.8	2.8	3.0
3.0	-2.8	2.8	3.0

At most one “extra” large residual occurs.

**Residuals versus fitted values:** If the model is correctly specified, the fitted values and residuals are uncorrelated in their sampling population. A scatterplot of  $Y - \hat{Y}$  against  $\hat{Y}$  should therefore show “no structure.”

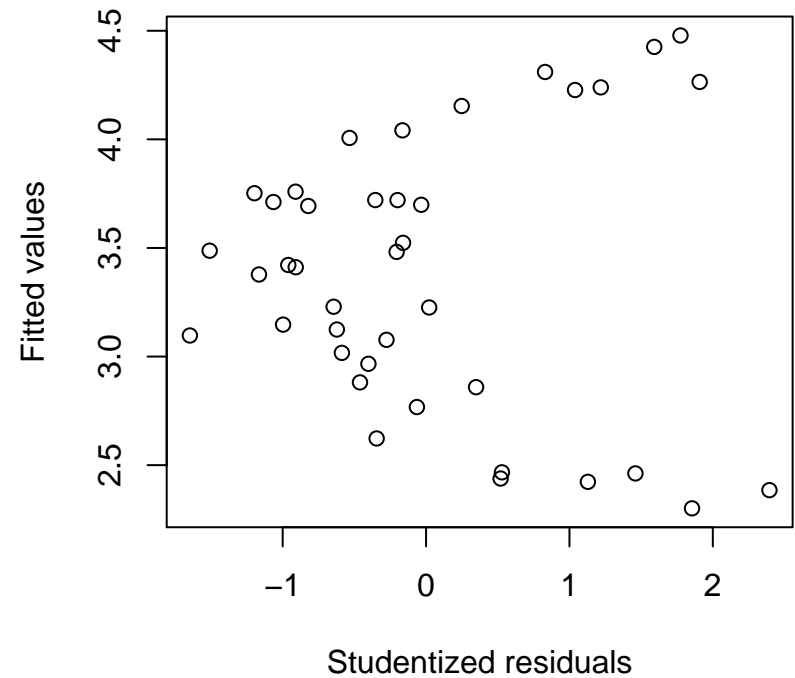
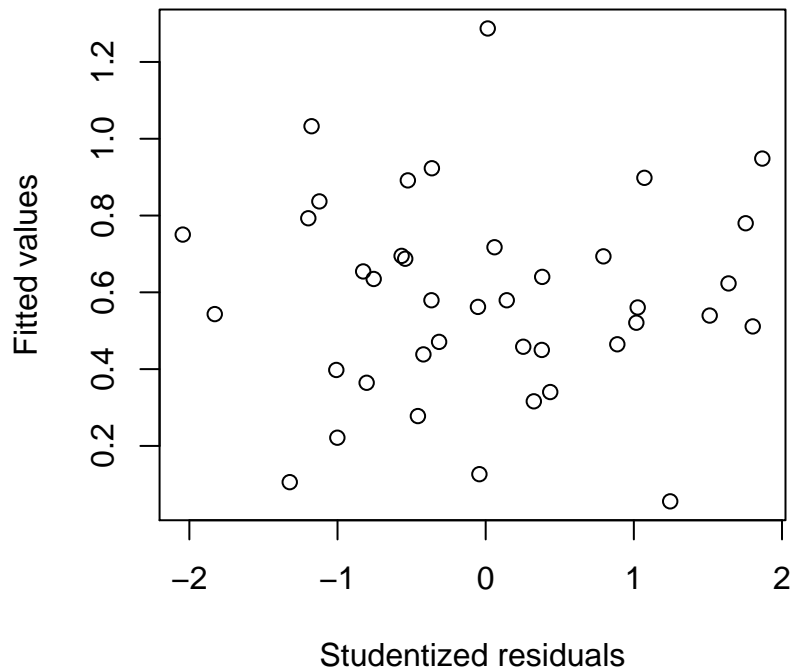
Regardless of the truth of the model specification, the residuals and fitted values have zero sample correlation in a particular least squares fit. Therefore we must look beyond linear trends in this plot to detect violations of the model assumptions.

*Example:* The following shows two “fitted values on residuals” diagnostic plots. The first is for the model

$$Y = \frac{1}{1 + (X_1 + X_2)^2} + \epsilon$$

and the second is for the model

$$Y = 1 + (X_1 + X_2)^2 + \epsilon.$$

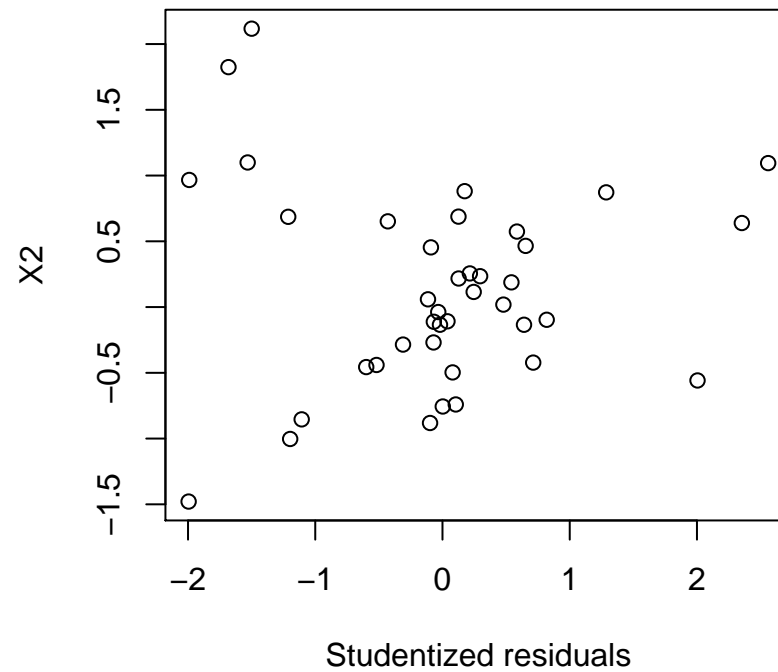


In the first case the diagnostic plot fails to detect a problem. In the second case we are alerted to the problem by the pattern of higher-order dependence between the fitted values and residuals.

**Residuals versus covariates:** Plots of the Studentized residuals against the individual covariate data may reveal model specification problems.

The following plot shows covariate  $X_2$  plotted against studentized residuals for the model

$$Y = 5X_1 + 3|X_2|\epsilon.$$



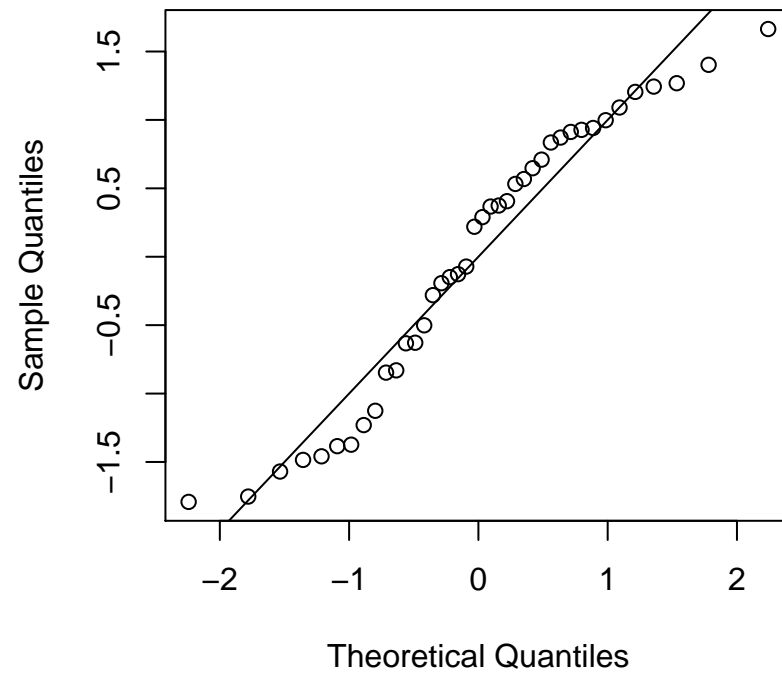
The tendency of points with large  $|X_2|$  values to have large residuals is weakly evident.

**Normal probability plot:** A normal probability plot of the residuals may reveal skewness or heavy tails in the errors.

The following normal probability plot of studentized residuals is based on data from the model

$$Y = X_1 + X_2 + \epsilon$$

where  $\epsilon$  is uniform on  $(-1, 1)$ .



**Other useful residual plots:** If there are “nuisance factors” such as the order in which the observations were made, or batches of observations that were collected in a common way (e.g. a common reagent batch was used), then residuals can be analyzed for patterns relating to the nuisance factor.

## Leverage and influence

“Leverage” is a measure of how strongly the data for case  $i$  determine the fitted value  $\hat{Y}_i$ .

Since  $\hat{Y} = PY$ , and

$$\hat{Y}_i = \sum_j P_{ij} Y_j,$$

it is natural to define the leverage for case  $i$  as  $P_{ii}$ , where  $P$  is the projection matrix onto  $\text{col}(X)$ .

The variance of the  $i^{\text{th}}$  residual is  $\sigma^2(1 - P_{ii})$ , when  $P_{ii}$  is close to 1, the fitted line will usually pass close to  $(X_i, Y_i)$ . This is undesirable, since when  $\epsilon_i$  happens to be large (by chance), the population regression surface will not come particularly close to  $Y_i$ , so the fitted surface should not either.

The average leverage is  $\text{trace}(P)/n = (p+1)/n$ . If the leverage for a particular case is two or more times greater than the average leverage, it is considered to have high leverage.

In simple linear regression, it is easy to show that

$$\text{var}(Y_i - \hat{\alpha} - \hat{\beta}X_i) = (n-1)\sigma^2/n - \sigma^2(X_i - \bar{X})^2 / \sum_j (X_j - \bar{X})^2.$$

This implies that when  $p = 1$ ,

$$P_{ii} = 1/n + (X_i - \bar{X})^2 / \sum_j (X_j - \bar{X})^2.$$

In general,

$$P_{ii} = X_i(X'X)^{-1}X_i' = X_i(X'X/n)^{-1}X_i'/n$$

where  $X_i$  is the  $i^{\text{th}}$  row of  $X$  (including the intercept).

Let  $\tilde{X}_i$  be row  $i$  of  $X$  without the intercept, let  $\tilde{\mu}$  be the sample mean of the  $\tilde{X}_i$ , and let  $\Sigma_X$  be the sample variance of the  $\tilde{X}_i$  (scaled by  $n$  rather than  $n - 1$ ). It is a fact that

$$X_i(X'X/n)^{-1}X_i' = (\tilde{X}_i - \tilde{\mu})\Sigma_X^{-1}(\tilde{X}_i - \tilde{\mu})' + 1$$

and therefore

$$P_{ii} = ((\tilde{X}_i - \tilde{\mu}_X)\Sigma_X^{-1}(\tilde{X}_i - \tilde{\mu}_X)' + 1) / n.$$

Note that this implies that  $P_{ii} \geq 1/n$ .

The next few slides will prove the result about  $P_{ii}$  given on the previous slide.

Let  $x = (1 \ x_1)'$  and  $y = (1 \ y_1)'$  be two vectors. Let  $\Sigma_x \in \mathcal{R}^{p \times p}$  be the sample covariance matrix and  $\mu \in \mathcal{R}^p$  be the sample mean of columns 2 through  $p+1$  of  $X$ .

We will show that the following two quantities differ by 1:

$$(x_1 - \mu)' \Sigma^{-1} (y_1 - \mu)$$

and

$$x' S^{-1} y,$$

where  $S = n^{-1} X' X$ .

To show this, let  $S_0$  be the lower right  $p \times p$  block of  $S$ , so that

$$\Sigma = S_0 - \mu\mu'.$$

It follows that

$$\Sigma^{-1} = S_0^{-1} + \frac{S_0^{-1}\mu\mu'S_0^{-1}}{1 - \mu'S_0^{-1}\mu}.$$

Now if we take  $(x_1 - \mu)' \Sigma^{-1} (y_1 - \mu)$  and do some algebra we get

$$\frac{x_1' (cS_0^{-1} + S_0^{-1} \mu \mu' S_0^{-1}) y_1 - \mu' S_0^{-1} (x_1 + y_1) + \mu' S_0^{-1} \mu}{c}$$

where  $c = 1 - \mu' S_0^{-1} \mu$ .

Now if we write

$$S = \begin{pmatrix} 1 & \mu' \\ \mu & S_0 \end{pmatrix}$$

and invert by blocks we get

$$S^{-1} = \begin{pmatrix} 1/c & -\mu' S_0^{-1} / c \\ -S_0^{-1} \mu / c & S_0^{-1} + S_0^{-1} \mu \mu' S_0^{-1} / c \end{pmatrix}.$$

So if we take  $x'S^{-1}y$  we get

$$1/c - \mu'S_0^{-1}(x_1 + y_1)/c + x_1'S_0^{-1}y_1 + x_1'S_0^{-1}\mu\mu'S_0^{-1}y_1/c.$$

which finally yields

$$x'S^{-1}y = (x - \mu)'\Sigma^{-1}(y - \mu) + 1.$$

## Influence

“Influence” measures the degree to which deletion of a case changes the fitted model.

This is different from leverage – a high leverage point has the potential to be influential, but if it happens to lie on the line determined by the other points, deletion of the case will have little effect.

The “deleted slope” for case  $i$  is the fitted slope vector that obtained upon deleting case  $i$ . The following identity allows the deleted slopes to be calculated efficiently

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{R_i}{1 - P_{ii}}(X'X)^{-1}X'_{i:},$$

where  $R_i$  is the  $i^{\text{th}}$  residual, and  $X_{i:}$  is row  $i$  of the design matrix.

The deleted fitted values  $\hat{Y}_{(i)}$  are

$$\hat{Y}_{(i)} = X\hat{\beta}_{(i)} = \hat{Y} - \frac{R_i}{1 - P_{ii}} X(X'X)^{-1} X'_{i:}$$

Influence can be measured by “Cook’s distance:”

$$\begin{aligned} D_i &\equiv \frac{1}{(p+1)\hat{\sigma}^2} (\hat{Y} - \hat{Y}_{(i)})' (\hat{Y} - \hat{Y}_{(i)}) \\ &= \frac{R_i^2}{(1 - P_{ii})^2 (p+1)\hat{\sigma}^2} X_{i:} (X'X)^{-1} X'_{i:} \\ &= \frac{P_{ii} R_i^{s2}}{(1 - P_{ii})(p+1)}, \end{aligned}$$

where  $R_i$  is the residual and  $R_i^s$  is the studentized residual. This will be large only if both the leverage  $P_{ii}$  is high, and the studentized residual is large.

Why is Cook's distance scaled by  $p + 1$  rather than  $n$ ? Since  $\hat{Y} - \hat{Y}_{(i)}$  lies in a  $p + 1$  dimensional subspace.

Cook's distance approximately captures the average squared change in fitted values due to deleting case  $i$ , in error variance units.

As a general rule,  $D_i$  values from  $1/2$  to  $1$  are high, and values greater than  $1$  are considered to be a problem.

## PRESS residuals

If case  $i$  is deleted and a prediction is made, we can compare the observed and predicted values to get the “prediction residual:”

$$R_{(i)} \equiv Y_i - \hat{Y}_{(i)i}.$$

A simple formula for the prediction residual is given by

$$\begin{aligned} R_{(i)} &= Y_i - X_i \hat{\beta}_{(i)} \\ &= Y_i - X_i (\hat{\beta} - R_i (X'X)^{-1} X_i / (1 - P_{ii})) \\ &= R_i / (1 - P_{ii}). \end{aligned}$$

The sum of squares of the prediction residuals is called “PRESS” (prediction error sum of squares). It is equivalent to using leave-one-out cross validation to estimate the generalization error rate.

## Transformations

If the residual diagnostics suggest that the linear model assumptions do not hold, it may be possible to continuously transform either  $Y$  or  $X$  so that the linear model becomes more consistent with the data.

## Variance stabilizing transformations

A common violation of the linear model assumptions is a “mean/variance relationship,” where  $EY_i$  and  $\text{var}(Y_i)$  are related.

Suppose that

$$\text{var } Y_i = g(EY_i)\sigma^2,$$

and let  $f(\cdot)$  be a transform to be applied to the  $Y_i$ . The goal is to find a transform such that the variances of the transformed responses are constant. Using a Taylor expansion,

$$f(Y_i) \approx f(EY_i) + f'(EY_i)(Y_i - EY_i).$$

Therefore

$$\text{var } f(Y_i) \approx f'(EY_i)^2 \text{var}(Y_i) = f'(EY_i)^2 g(EY_i) \sigma^2.$$

The goal is to find  $f$  such that  $f' = 1/\sqrt{g}$ .

*Example:* Suppose  $g(z) = z^\lambda$ . This includes the “Poisson regression” case  $\lambda = 1$ , where the variance is proportional to the mean, and the case  $\lambda = 2$  where the standard deviation is proportional to the mean.

When  $\lambda = 1$ ,  $f$  solves  $f'(z) = 1/\sqrt{z}$ , so  $f$  is the square root function.

When  $\lambda = 2$ ,  $f$  solves  $f'(z) = 1/z$ , so  $f$  is the logarithm function.

## *Log/log regression*

Suppose we fit a simple linear regression of the form

$$E(\log(Y)|\log(X)) = \alpha + \beta \log(X).$$

Suppose the logarithms are base 10. Let  $X_z = X \cdot 10^z$ . Under the model,

$$E(\log(Y)|X_z) - E(\log(Y)|X) = \beta z$$

Using the crude approximation  $\log E(Y|X) \approx E(\log(Y)|X)$ , we conclude  $E(Y|X)$  is approximately scaled by a factor of  $10^{\beta z}$  when  $X$  is scaled by a factor of  $10^z$ . This holds for relatively small values of  $z$  where the “crude approximation” holds.

Thus in a log/log model, we may say that a  $f\%$  change in  $X$  is approximately associated with a  $f^{\beta}\%$  change in the expected response.

### *Maximum likelihood estimation of a transformation*

The Box-Cox family of transforms is

$$y \rightarrow \frac{y^\lambda - 1}{\lambda},$$

which makes sense only when all  $Y_i$  are positive.

The Box-Cox family includes the identity ( $\lambda = 1$ ), all power transformations such as the square root ( $\lambda = 1/2$ ) and reciprocal ( $\lambda = -1$ ), and the logarithm in the limiting case  $\lambda \rightarrow 0$ .

Suppose we assume that for some value of  $\lambda$ , the transformed data follow a linear model with Gaussian errors. We can then set out to estimate  $\lambda$ .

The joint log-likelihood of the transformed data is

$$-\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_i (Y_i^{(\lambda)} - X_i' \beta)^2.$$

Next we transform this back to a likelihood in terms of  $Y_i = g_\lambda^{-1}(Y_i^{(\lambda)})$ . This joint log-likelihood is

$$-\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_i (g_\lambda(Y_i) - X_i' \beta)^2 + \sum_i \log J_i$$

where the Jacobian is

$$\log J_i = \log g_\lambda'(Y_i) = (\lambda - 1) \log Y_i.$$

The joint log likelihood for the  $Y_i$  is

$$-\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_i (g_\lambda(Y_i) - X_i' \beta)^2 + (\lambda - 1) \sum_i \log Y_i.$$

This likelihood is maximized with respect to  $\lambda$ ,  $\beta$ , and  $\sigma^2$  to identify the MLE.

To do the maximization, let  $Y^{(\lambda)} \equiv g_\lambda(Y)$  denote the transformed observed responses, and let  $\hat{Y}^{(\lambda)}$  denote the fitted values from regressing  $Y^{(\lambda)}$  on  $X$ . Since  $\sigma^2$  does not appear in the Jacobian,

$$\hat{\sigma}_\lambda^2 \equiv n^{-1} \|Y^{(\lambda)} - \hat{Y}^{(\lambda)}\|^2$$

will be the maximizing value of  $\sigma^2$ . Therefore the MLE of  $\beta$  and  $\lambda$  will maximize

$$-\frac{n}{2} \log \hat{\sigma}_\lambda^2 + (\lambda - 1) \sum_i \log Y_i.$$

## Collinearity Diagnostics

Collinearity inflates the sampling variances of covariate effect estimates.

To understand the effect of collinearity on  $\text{var}\hat{\beta}_j$ , reorder the columns and partition the design matrix  $X$  as

$$X = ( X_j \mid X_0 ) = ( X_j - X_j^\perp + X_j^\perp \mid X_0 )$$

where  $X_0$  is the  $n \times p$  matrix consisting of all columns in  $X$  except  $X_j$ , and  $X_j^\perp$  is the projection of  $X_j$  onto  $\text{col}(X_0)^\perp$ . Therefore

$$H \equiv X'X = \left( \begin{array}{c|c} X_j'X_j & (X_j - X_j^\perp)'X_0 \\ \hline X_0'(X_j - X_j^\perp) & X_0'X_0 \end{array} \right).$$

$\text{var}\hat{\beta}_j = \sigma^2 H_{11}^{-1}$ , so we want a simple expression for  $H_{11}^{-1}$ .

A symmetric block matrix can be inverted using:

$$\begin{pmatrix} A & B \\ B' & C \end{pmatrix}^{-1} = \begin{pmatrix} S^{-1} & -S^{-1}BC^{-1} \\ -C^{-1}B'S^{-1} & C^{-1} + C^{-1}B'S^{-1}BC^{-1} \end{pmatrix},$$

where

$$S = A - BC^{-1}B'.$$

Therefore

$$H_{1,1}^{-1} = \frac{1}{\|X_j\|^2 - (X_j - X_j^\perp)'P_0(X_j - X_j^\perp)},$$

where  $P_0 = X_0(X_0'X_0)^{-1}X_0$  is the projection matrix onto  $\text{col}(X_0)$ .

Since  $X_j - X_j^\perp \in \text{col}(X_0)$ , we can write

$$H_{1,1}^{-1} = \frac{1}{\|X_j\|^2 - \|X_j - X_j^\perp\|^2},$$

and since  $X_j^{\perp'}(X_j - X_j^\perp) = 0$ , it follows that

$$\|X_j\|^2 = \|X_j - X_j^\perp + X_j^\perp\|^2 = \|X_j - X_j^\perp\|^2 + \|X_j^\perp\|^2,$$

so

$$H_{1,1}^{-1} = \frac{1}{\|X_j^\perp\|^2}.$$

This makes sense, since smaller values of  $\|X_j^\perp\|^2$  correspond to greater collinearity.

Let  $R_{jx}^2$  be the coefficient of determination (multiple  $R^2$ ) for the regression of  $X_j$  on the other covariates.

$$R_{jx}^2 = 1 - \frac{\|X_j - (X_j - X_j^\perp)\|^2}{\|X_j - \bar{X}_j\|^2} = 1 - \frac{\|X_j^\perp\|^2}{\|X_j - \bar{X}_j\|^2}.$$

Combining the two equations yields

$$H_{11}^{-1} = \frac{1}{\|X_j - \bar{X}_j\|^2} \cdot \frac{1}{1 - R_{jx}^2}.$$

The two factors in this expression reflect two different sources of variance of  $\hat{\beta}_j$ :

- $1/\|X_j - \bar{X}_j\|^2 = 1/((n-1)\widehat{\text{var}}(X_j))$  reflects the scaling of  $X_j$
- The “variance inflation factor” (VIF)  $1/(1 - R_{jx}^2)$  is scale-free. It is always greater than or equal to 1, and is equal to 1 only if  $X_j$  is orthogonal to the other covariates. Large values of the VIF indicate that parameter estimation is strongly affected by collinearity.