Regression diagnostics

Kerby Shedden

Department of Statistics, University of Michigan

October 31, 2021

Motivation

When working with a linear model with design matrix X, the conventional linear model is based on the following conditions:

$$E[y|X] \in \mathrm{col}(X) \quad \text{and} \quad \mathrm{var}[y|X] = \sigma^2 I_{n\times n}.$$

Unbiasedness and consistency of least squares point estimates depend on the first condition approximately holding. Least squares inferences depend on both of the above conditions approximately holding.

Inferences for small sample sizes may also depend on the distribution of $y - E[y|X]$ being approximately multivariate Gaussian, but for moderate or large sample sizes this condition is not critical.

Regression diagnostics for linear models are approaches for assessing how well a particular data set fits one or both of these conditions.

Residuals

Linear models can be expressed in two equivalent ways:

- Focus only on moments:

$$E[y|X] \in \mathrm{col}(X) \quad \text{and} \quad \mathrm{var}[y|X] = \sigma^2 I.$$

- Use a "generative model" such as an additive error model of the form $y = X\beta + \epsilon$, where ϵ is random with $E[\epsilon|X] = 0$, and $\mathrm{cov}[\epsilon|X] \propto I_{n\times n}$.

Since the residuals can be viewed as predictions of the errors, it turns out that regression model diagnostics can often be developed using the residuals.

Recall that the residuals can be expressed

$$r \equiv (I - P)y$$

where P is the projection onto $\mathrm{col}(X)$ and $y \in \mathcal{R}^n$ is the vector of observed responses.

Residuals

The residuals have two key mathematical properties regardless of the correctness 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 are orthogonal (they have zero sample covariance):

$$\begin{aligned} \widehat{\text{cov}}(r, \hat{y}|X) &\propto (r - \bar{r})'\hat{y} \\ &= r'\hat{y} \\ &= y'(I - P)Py \\ &= 0. \end{aligned}$$

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

Residuals

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

- The expected value of each residual is zero:

$$\begin{aligned} E[r|X] &= (I - P)E[y|X] \\ &= 0 \in \mathcal{R}^n. \end{aligned}$$

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

$$\begin{aligned} \mathrm{cov}(r, \hat{y}|X) &= E[r\hat{y}'] \\ &= (I - P)\mathrm{cov}(y|X)P \\ &= \sigma^2(I - P)P \\ &= 0 \in \mathcal{R}^{n \times n}. \end{aligned}$$

Residuals

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

$$\begin{aligned} \operatorname{cov}(r|X) &= (I-P)E[yy'](I-P) \\ &= (I-P)\left(X\beta\beta'X' + \sigma^2 I\right)(I-P) \\ &= \sigma^2(I-P). \end{aligned}$$

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 zero mean and unit variance.

Residuals

The standardized residuals are crudely standardized using a plug-in logic. The standardized "errors" are

$$\frac{y_i - E[y|\mathbf{x} = x_i]}{\sigma},$$

which have zero mean and unit variance (exactly).

If we plug-in $\hat{y}_i$ for $E[y|\mathbf{x} = x_i]$ and $\hat{\sigma}$ for σ, then we get the standardized residual.

However we know that the actual variance of $y_i - \hat{y}_i$ is $\sigma^2(I - P)_{ii}$, so it is more precise to scale by $\sigma\sqrt{(I - P)_{ii}}$ rather than scaling by σ.

Since we plug in the estimate $\hat{\sigma}$ for the population parameter σ, even the Studentized residual does not have variance exactly equal to 1.

External standardization of residuals

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

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

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

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

Outliers and masking

In some settings, residuals can be used to identify "outliers". However, in a small data set, a large outlier will increase the value of $\hat{\sigma}$, and hence may mask itself.

Externally Studentized residuals solve the problem of a single large outlier masking itself. But masking may still occur if multiple large outliers are present.

Outliers and masking

If multiple large outliers may be present we may use alternate estimates of the scale parameter σ:

- Interquartile range (IQR): this is the difference between the 75^{th} percentile and the 25^{th} percentile of the distribution or data. The IQR of the standard normal distribution is 1.35, so $\text{IQR}/1.35$ can be used to estimate σ.
- Median Absolute Deviation (MAD): this is the median value of the absolute deviations from the median of the distribution or data, i.e. $\text{median}(|Z - \text{median}(Z)|)$. The MAD of the standard normal distribution is 0.65, so $\text{MAD}/0.65$ can be used to estimate σ.

These alternative estimates of σ can be used in place of the usual $\hat{\sigma}$ for standardizing or Studentizing residuals.

Leverage

Leverage is a measure of how strongly the data for case i determines the fitted value $\hat{y}_i$.

Since $\hat{y} = Py$, where P is the projection matrix onto $\mathrm{col}(X)$, then

$$\hat{y}_i = \sum_j P_{ij} y_j.$$

It is natural to define the leverage for case i as P_{ii},

This is related to the fact that the variance of the i^{th} residual is $\sigma^2(1 - P_{ii})$. Since the residuals have mean zero, when P_{ii} is close to 1, the residual will likely be close to zero. Thus the least squares fit will tend to pass closer to high leverage points than low leverage points.

Leverage

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

In simple linear regression, we showed earlier that

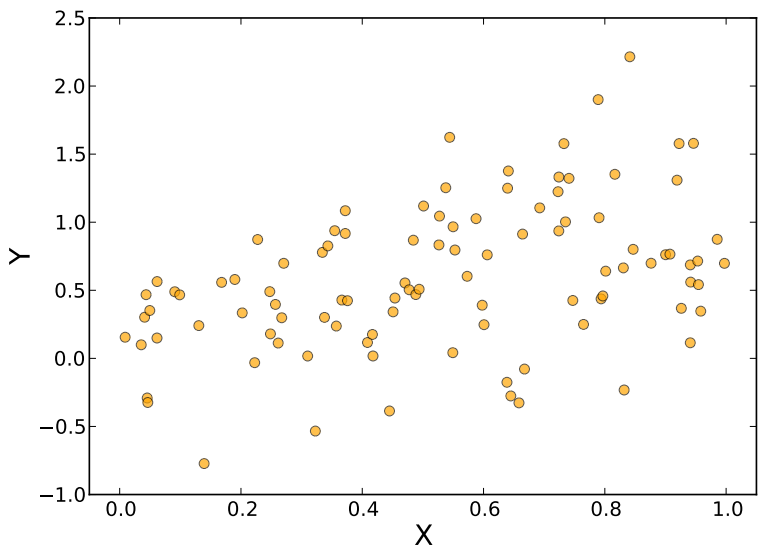
$$\mathrm{var}(y_i - \hat{\alpha} - \hat{\beta}x_i) = (n-1)\sigma^2/n - \sigma^2(x_i - \bar{x})^2/\sum_j (x_j - \bar{x})^2.$$

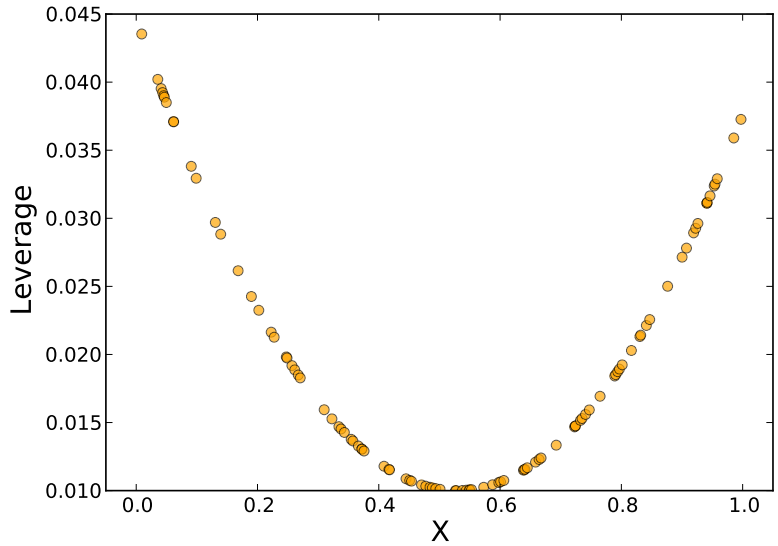
Since $\mathrm{var}(r_i) = \sigma^2(1 - P_i i)$, this implies that when $p = 1$,

$$P_{ii} = 1/n + (x_i - \bar{x})^2/\sum_j (x_j - \bar{x})^2.$$

Leverage

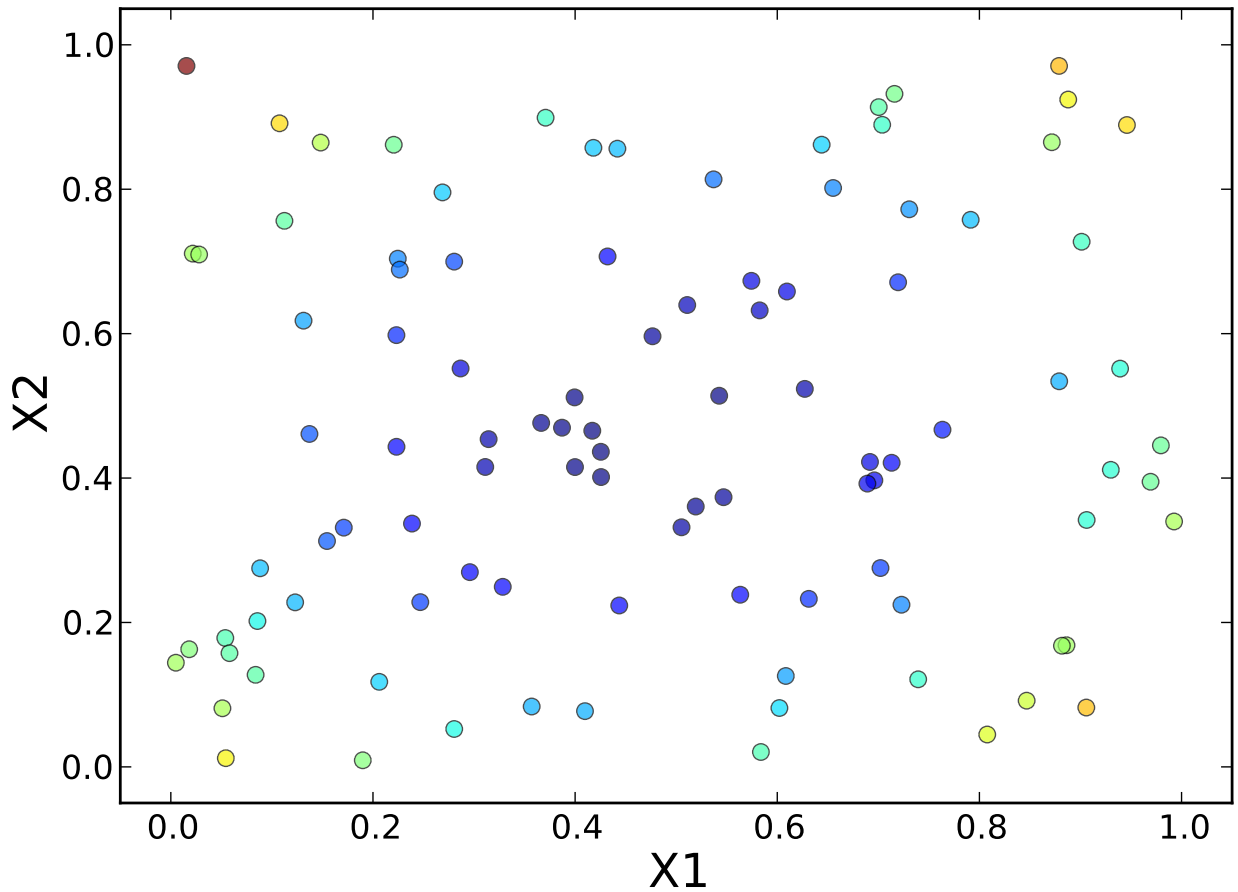
Leverage values in a simple linear regression:





Leverage

Leverage values in a linear regression with two independent variables:



Projection matrices and Mahalanobis distances

Suppose that the design matrix X has a leading column of 1's. The projection onto $\mathrm{col}(X)$ is unchanged if we transform X to XB, where $B \in \mathcal{R}^{p+1 \times p+1}$ is invertible. Thus, we can take $\tilde{X} \in \mathcal{R}^{n \times p+1}$ to be a matrix with leading column identically equal to 1, and for $j > 1$ the j^{th} column of $\tilde{X}$ is the centered version of the j^{th} column of X.

Since

$$P \equiv \mathrm{proj}(\mathrm{col}(X)) = \mathrm{proj}(\mathrm{col}(\tilde{X})) = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}',$$

it follows that

$$P_{ij} = \tilde{x}_i(\tilde{X}'\tilde{X})^{-1}\tilde{x}_j'$$

where $\tilde{x}_i$ is row i of $\tilde{X}$.

Projection matrices and Mahalanobis distances

Let $\Sigma_X \in \mathcal{R}^{p \times p}$ denote the covariance matrix of columns 2 through $p+1$ of X (or of $\tilde{X}$). We can block decompose as follows:

$$\tilde{X}'\tilde{X}/n = \begin{pmatrix} 1_{1\times 1} & 0_{1\times p} \\ 0_{p\times 1} & \Sigma_X \end{pmatrix}.$$

We can write $\tilde{x}_i = (1, \ x_i^* - \mu_X)$, where $\mu_X \in \mathcal{R}^p$ is the mean of columns 2 through $p+1$ of X and x_i^* contains elements 2 through $p+1$ of x_i.

This gives us

$$P_{ij} = n^{-1}(1 + (x_i^* - \mu_X)\Sigma_X^{-1}(x_j^{*\prime} - \mu_X)).$$

Since Σ_X^{-1} is positive definite, this implies that $P_{ii} \geq 1/n$.

Leverage and Mahalanobis distances

The expression

$$(x_i^* - \mu_X)\Sigma_X^{-1}(x_i^* - \mu_X)'$$

is the Mahalanobis distance between x_i^* and μ_X. It is equivalent to decorrelating the x_i^* to $z_i = R^{-T}x_i^*$ (e.g. where $\Sigma_X = R'R$) and then taking the squared Euclidean distance $\|z_i - R^{-T}\mu_X\|^2$.

Thus there is a direct relationship between the Mahalanobis distance of a point relative to the center of the covariate set (μ_X), and its leverage.

Observations with covariate values x_i^* in the tails of the distribution of all $\{x_i^*\}$, as assessed with Mahalanobis distance, have the highest leverage (but no larger than 1).

Observations with covariate values close to the centroid μ_X of all $\{x_i^*\}$ have the least leverage (but no smaller than $1/n$).

Influence

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

We will see that this is different from leverage – a high leverage point has the potential to be influential, but is not always influential.

The deleted slope for case i is the fitted slope vector that is 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^{th} residual, and x_i is row i of the design matrix.

Influence

The vector of all deleted fitted values $\hat{y}_{(i)}$ are

$$\hat{y}_{(i)} = X\hat{\beta}_{(i)} = \hat{y} - \frac{r_i}{1 - P_{ii}} X(X'X)^{-1}X'.$$

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.

Influence

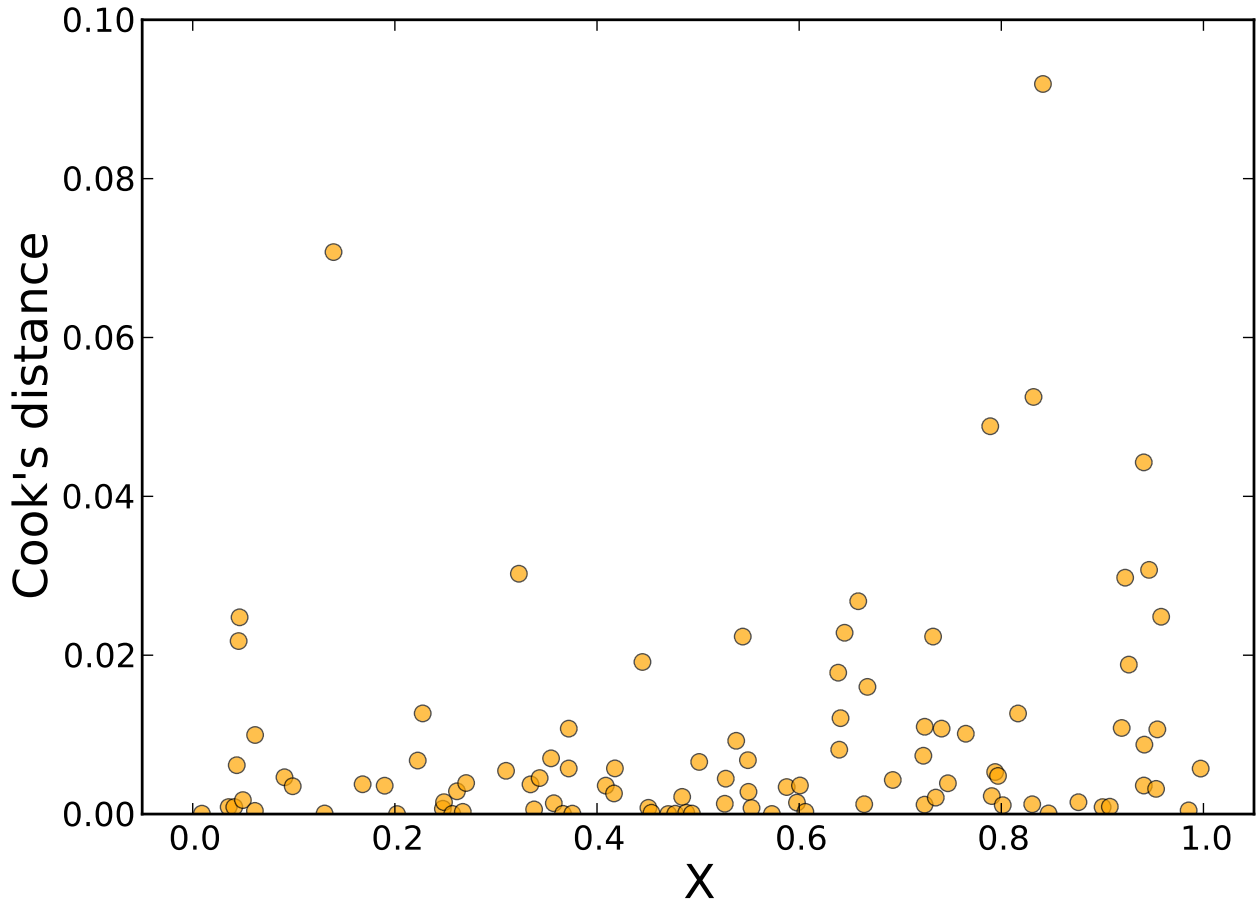
Cook's distance approximately captures the average squared change in fitted values due to deleting case i, in error variance units.

Cook's distance is large only if both the leverage P_{ii} is high, and the studentized residual for the i^{th} case is large.

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

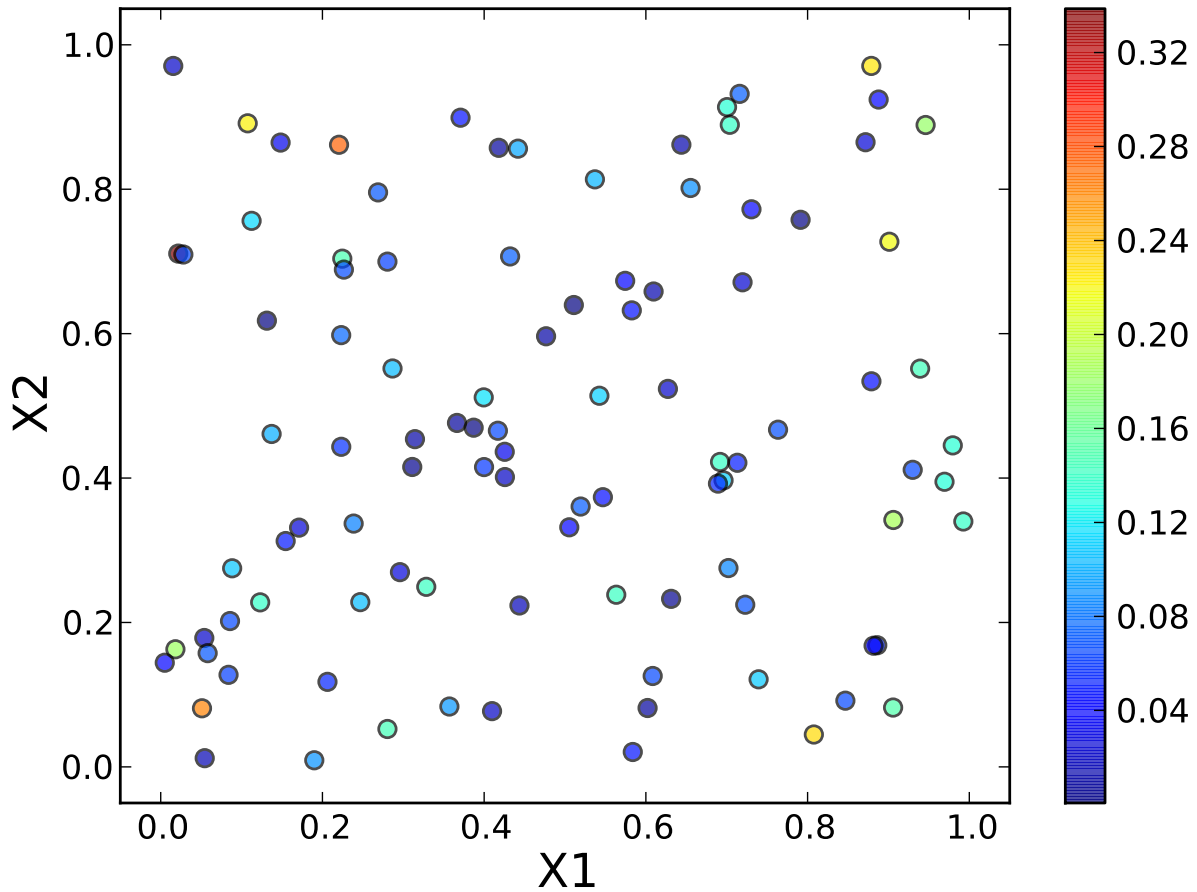
Influence

Cook's distances in a simple linear regression:



Influence

Cook's distances in a linear regression with two variables:



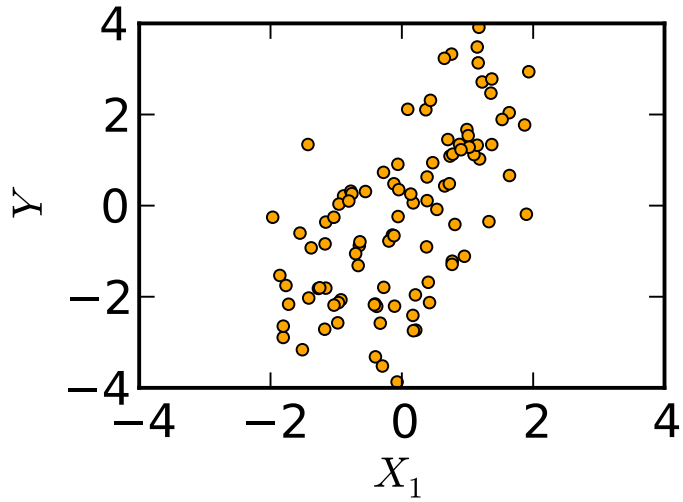
Regression graphics

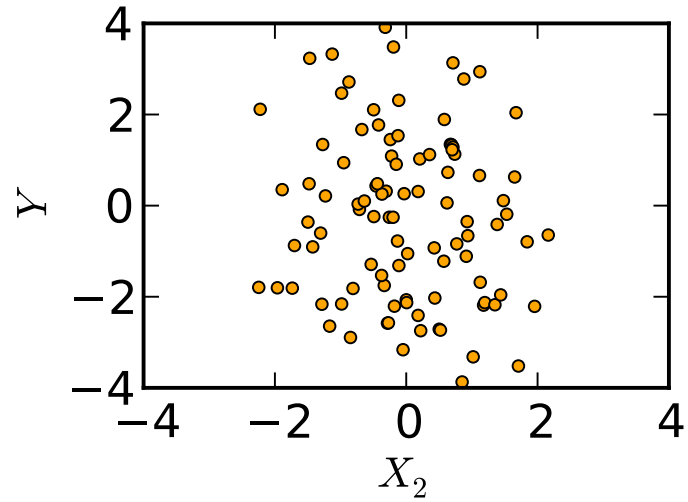
Quite a few graphical techniques have been proposed to aid in visualizing regression relationships. We will discuss the following plots:

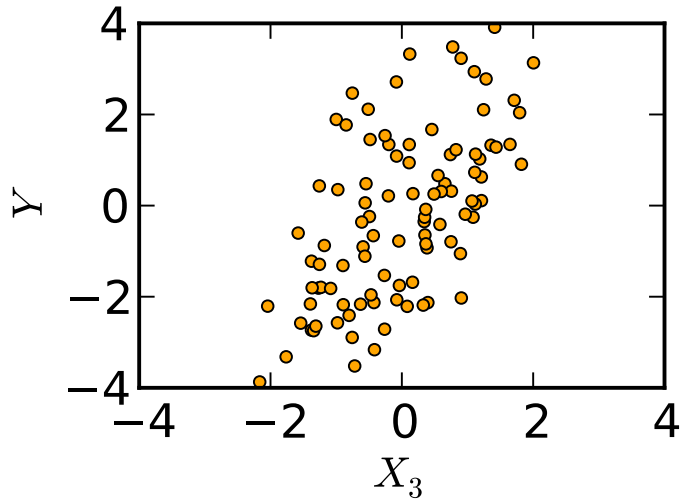
1. Scatterplots of y against individual covariates x_j
2. Scatterplots of two covariates x_j, x_k against each other
3. Residuals versus fitted values plot
4. Added variable plots
5. Partial residual plots
6. Residual quantile plots

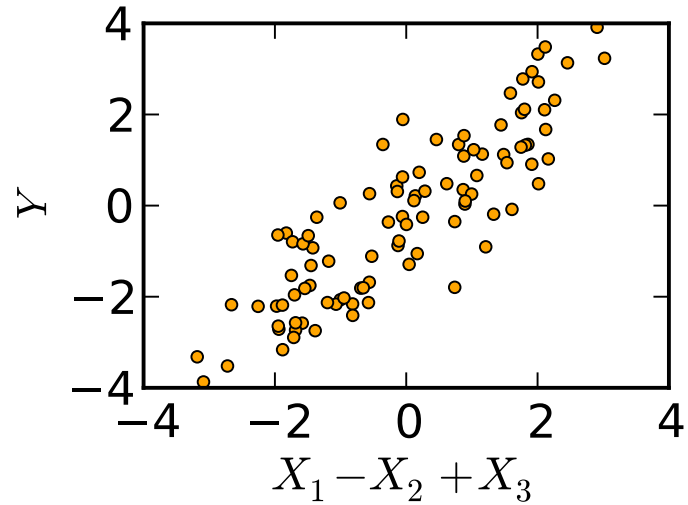
Scatterplots of y against individual covariates

$$E[y|x] = x_1 - x_2 + x_3,\ \mathrm{var}[y|x] = 1,\ \mathrm{var}(x_j) = 1,\ \mathrm{cor}(x_j, x_k) = 0.3$$



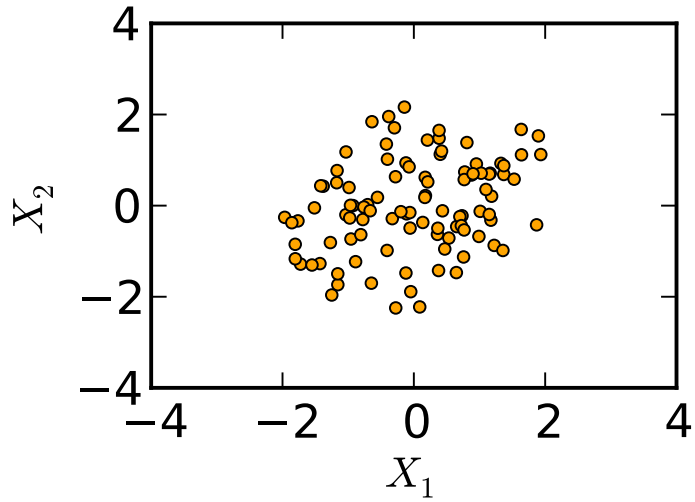


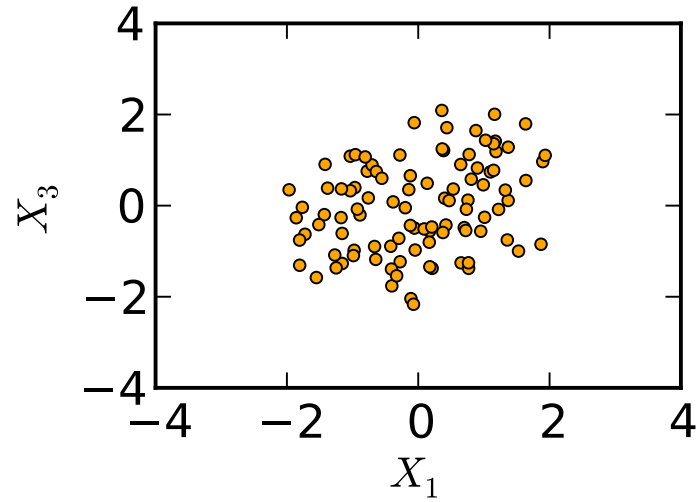


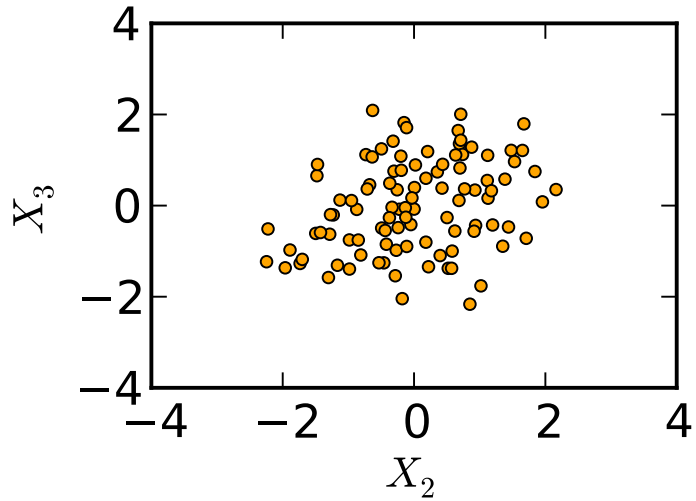


Scatterplots of covariates against each other

$E[y|x] = x_1 - x_2 + x_3$, $\text{var}[y|x] = 1$, $\text{var}(x_j) = 1$, $\text{cor}(x_j, x_k) = 0.3$

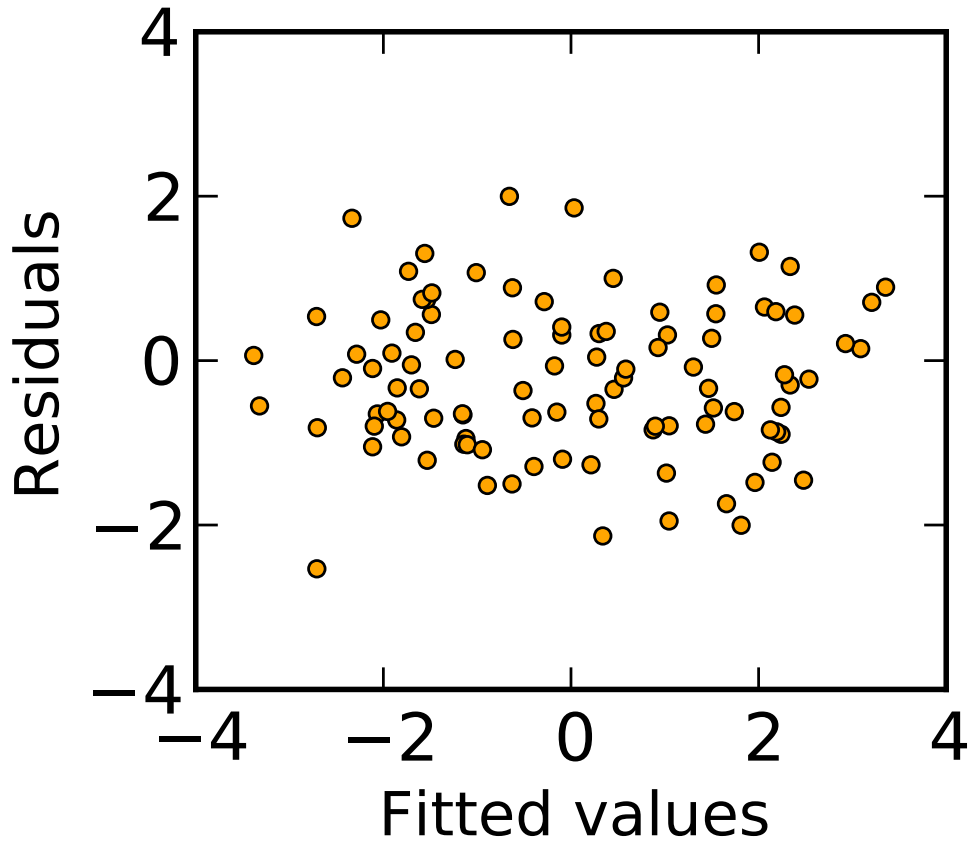






Residuals against fitted values plot

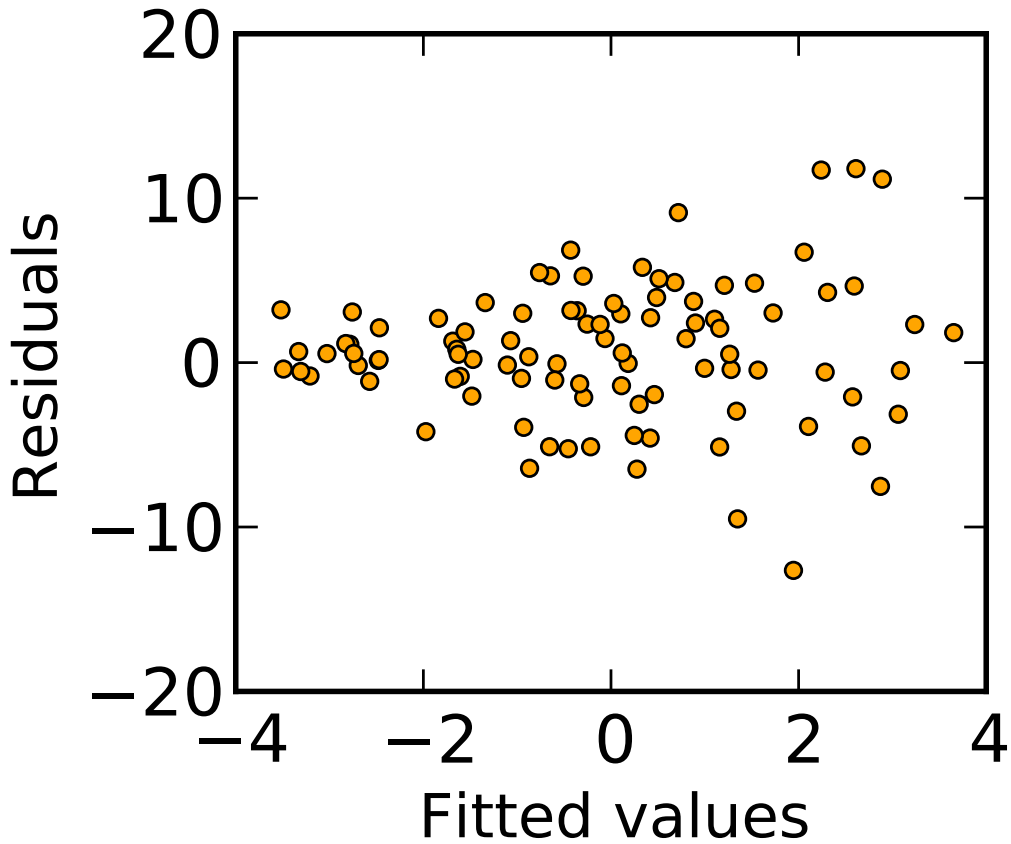
$E[y|x] = x_1 - x_2 + x_3$, $\mathrm{var}[y|x] = 1$, $\mathrm{var}(x_j) = 1$, $\mathrm{cor}(x_j, x_k) = 0.3$



Residuals against fitted values plots

Heteroscedastic errors:

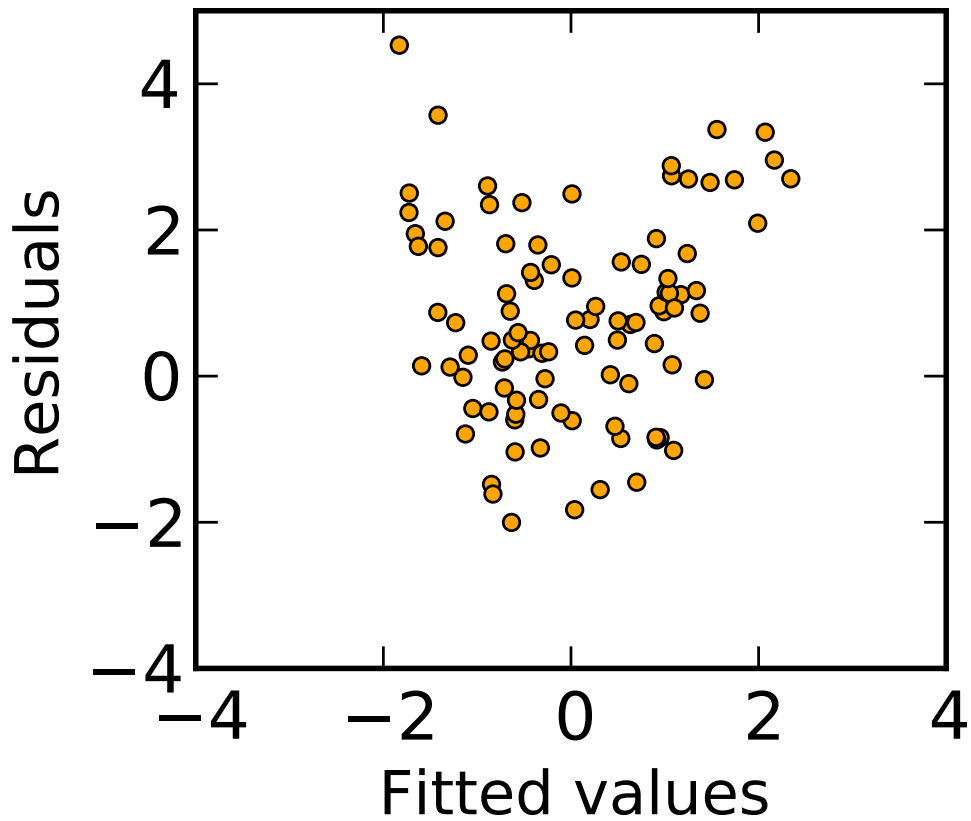
$E[y|x] = x_1 + x_3$, $\mathrm{var}[y|x] = 4 + x_1 + x_3$, $\mathrm{var}(x_j) = 1$, $\mathrm{cor}(x_j, x_k) = 0.3$



Residuals against fitted values plots

Nonlinear mean structure:

$E[y|x] = x_1^2$, $\mathrm{var}[y|x] = 1$, $\mathrm{var}(x_j) = 1$, $\mathrm{cor}(x_j, x_k) = 0.3$



Added variable plots

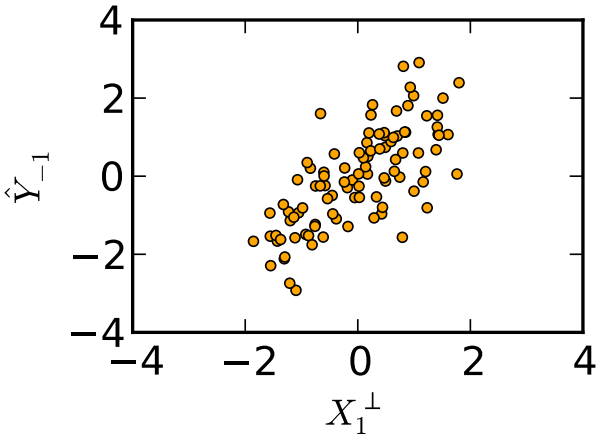
Suppose P_{-j} is the projection onto the span of all covariates except x_j, and define $\hat{y}_{-j} = P_{-j}y$, $x_j^\star = P_{-j}x_j$. The added variable plot is a scatterplot of $y - \hat{y}_{-j}$ against $x - x_j^\star$.

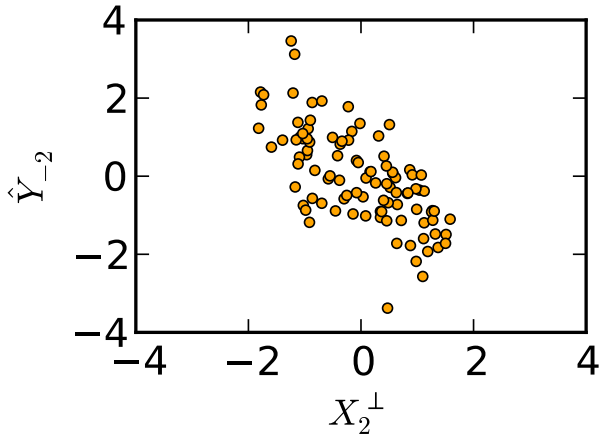
The squared correlation coefficient of the points in the added variable plot is the partial R^2 for variable j.

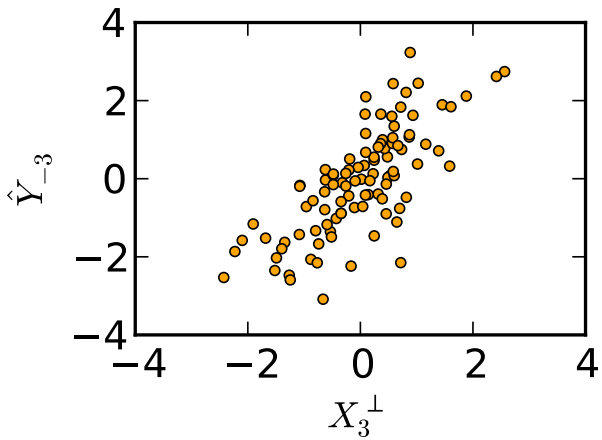
Added variable plots are also called partial regression plots.

Added variable plots

$E[y|x] = x_1 - x_2 + x_3$, $\mathrm{var}[y|x] = 1$, $\mathrm{var}(x_j) = 1$, $\mathrm{cor}(x_j, x_k) = 0.3$







Partial residual plot

Suppose we fit the model

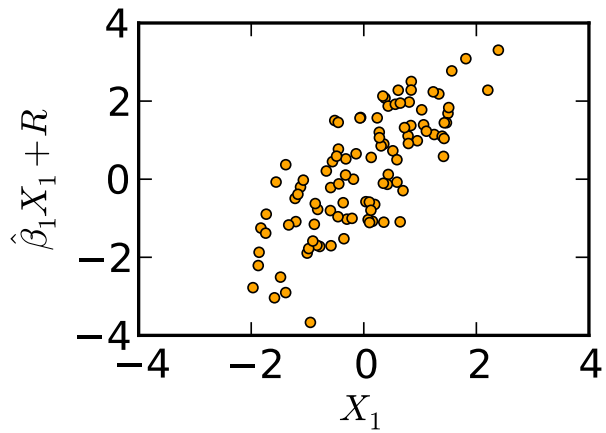
$$\hat{y}_i = \hat{\beta}' x_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots \hat{\beta}_p x_{ip}.$$

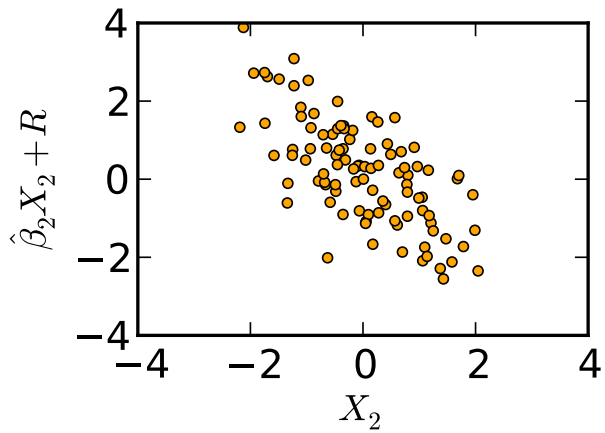
The partial residual plot for covariate j is a plot of $\hat{\beta}_j x_{ij} + r_i$ against x_{ij}, where r_i is the residual.

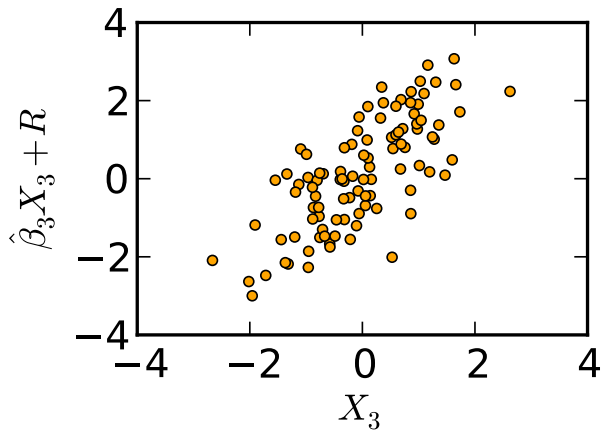
The partial residual plot attempts to show how covariate j is related to y, if we control for the effects of all other covariates.

Partial residual plot

$E[y|x] = x_1 - x_2 + x_3$, $\mathrm{var}[y|x] = 1$, $\mathrm{var}(x_j) = 1$, $\mathrm{cor}(x_j, x_k) = 0.3$



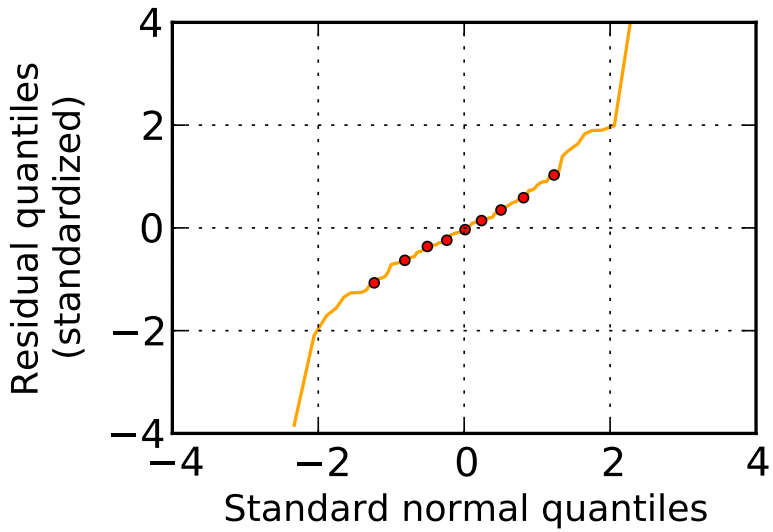




Residual quantile plots

$E[y|x] = x_1 - x_2 + x_3$, $\mathrm{var}[y|x] = 1$, $\mathrm{var}(x_j) = 1$, $\mathrm{cor}(x_j, x_k) = 0.3$

t_4 distributed errors



Transformations

As noted above, the linear model imposes two main constrains on the population that is under study. Specifically, the conditional mean function should be linear, and the conditional variance function should be constant.

If it appears that $E[y|\mathbf{x} = x]$ is not linear in x, or that $\mathrm{Var}[y|\mathbf{x} = x]$ is not constant in x, 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

Many populations encountered in practice exhibit a mean/variance relationship, where $E[y_i]$ and $\mathrm{var}[y_i]$ are related.

Suppose that

$$\mathrm{var}[y_i] = g(E[y_i])\sigma^2,$$

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,

$$f(y_i) \approx f(E[y_i]) + f'(E[y_i])(y_i - E[y_i]).$$

Variance stabilizing transformations

Therefore

$$\mathrm{var}[f(y_i)] \approx f'(E[y_i])^2 \cdot \mathrm{var}[y_i] = f'(E[y_i])^2 \cdot g(E[y_i])\sigma^2.$$

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

$$E[\log(y) \mid \log(x)] = \alpha + \beta \log(x).$$

$$E[\log(y) \mid \mathbf{x} = x + 1] - E[\log(y) \mid \mathbf{x} = x] = \beta$$

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 e^{β} when X is scaled by a factor of e.

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 data transformation

The Box-Cox family of transforms is

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

which makes sense only when all y is 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 $\lambda \to 0$.

Maximum likelihood estimation of a data transformation

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

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

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

Maximum likelihood estimation of a data transformation

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 λ, β, and σ^2 to identify the MLE.

Maximum likelihood estimation of a data transformation

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 σ^2 does not appear in the Jacobian,

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

will be the maximizing value of σ^2. Therefore the MLE of β and λ will maximize

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

Collinearity Diagnostics

Collinearity inflates the sampling variances of covariate effect estimates.

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

$$X = \left(\; x_j \;\middle|\; X_0 \; \right) = \left(\; x_j - x_j^{\perp} + x_j^{\perp} \;\middle|\; X_0 \; \right)$$

where x_j is column j of X, 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 $\mathrm{col}(X_0)^{\perp}$. Therefore

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

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

Collinearity Diagnostics

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 $\mathrm{col}(X_0)$.

Collinearity Diagnostics

Since $x_j - x_j^{\perp} \in \mathrm{col}(X_0)$, 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.

Collinearity Diagnostics

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

$$R^2_{jx} = 1 - \frac{\|x_j - (x_j - x_j^{\perp})\|^2}{\|x_j - \bar{x}_j\|^2} = 1 - \frac{\|x_j^{\perp}\|^2}{\|x_j - \bar{x}_j\|^2}.$$

Combining the two equations yields

$$H_{11}^{-1} = \frac{1}{\|x_j - \bar{x}_j\|^2} \cdot \frac{1}{1 - R^2_{jx}}.$$

Collinearity Diagnostics

The two factors in the expression

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

reflect two different sources of variance of $\hat{\beta}_j$:

- $1/\|x_j - \bar{x}_j\|^2 = 1/\left((n-1)\widehat{\mathrm{var}}(x_j)\right)$ reflects the scaling of x_j and the sample size n.
- The variance inflation factor (VIF) $1/(1 - R_{jx}^2)$ is scale-free and sample size-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.