Regression

Steven F. Arnold
Professor of Statistics
Penn State University

Regression is the most commonly used statistical technique.

It is primarily concerned with fitting models to data. It is often not possible to draw strong inferences from regression as other techniques such as analysis of variance, because typically regression is based on an observational study rather than a randomized designed study. I think of regression as a procedure for separating the "signal" from the "noise".

Some Background

Some expectations

Let X be a random variable. Then the expectation of X is called the *mean* of X. If X is a random variable with mean μ, then the *variance* of X is defined by

$$\sigma^2 = var\,(X) = E\,(X-\mu)^2 = EX^2 - \mu^2$$

The *standard deviation* of X is the square root of the variance.

If X and Y are random variables with means μ and ν, then the *covariance* X and Y is defined by

$$cov\,(X,Y) = E\,(X-\mu)\,(Y-\nu) = EXY - \mu\nu$$

Although we shall not use correlation today, the *correlation coefficient* $\rho\,(X,Y)$ of X and is defined by

$$\rho(X,Y) = \frac{cov\,(X,Y)}{\sqrt{var\,(X)\,var\,(Y)}}$$

Some properties of expectation are the following

$$E\,(aX+b) = aEX + b,$$

$$
\begin{aligned}
var\left(aX+b\right) &= a^2 var\left(X\right) \\
E\left(aX+bY+c\right) &= aEX+bEY+c \\
var\left(aX+bY+c\right) &= a^2 var\left(X\right)+b^2 var\left(Y\right)+2ab cov\left(X,Y\right)
\end{aligned}
$$

Random vectors, mean vectors and covariancematrice

Let $Y_1, ...Y_n$ be random variables. Then

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

is a p=dimesional *random vector.* Then the *mean vector* μ =$E\mathbf{Y}$ and *covariance matrix* $\mathbf{\Sigma}$ =$cov\left(\mathbf{Y}\right)$ are defined by

$$\mu=\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, \mathbf{\Sigma}=\begin{pmatrix} \Sigma_{11} & \cdots & \Sigma_{1n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n1} & \cdots & \Sigma_{nn} \end{pmatrix}$$

where

$$
\begin{aligned}
\mu_i = EY_i,\ \Sigma_{ii} = var\left(Y_i\right), \\
\Sigma_{ij} = cov\left(Y_i, Y_j\right),\ i \neq j
\end{aligned}
$$

Then it can be shown that

$$
\begin{aligned}
E\left(\mathbf{AY}+\mathbf{b}\right) &= \mathbf{A}E\mathbf{Y}+\mathbf{b}, \\
cov\left(\mathbf{AY}+\mathbf{b}\right) &= \mathbf{A}cov\left(\mathbf{Y}\right)\mathbf{A}'.
\end{aligned}
$$

which is the basic result used in regression.

Note that the covariance matrix is a symmetric matrix. Further,

$$0 \leq var\left(\mathbf{a}'\mathbf{Y}\right) = \mathbf{a}'cov\left(\mathbf{Y}\right)\mathbf{a}$$

which implies that the covariance matrix is non-negative definite, which means there is a matrix $\mathbf{B}$ such that

$$\mathbf{BB}' = cov\left(\mathbf{Y}\right)$$

Such a matrix $\mathbf{B}$ is called a *square root* of the covariance matrix. Actually there are several such matrices. One of the most useful and easy to find in computer software is the Cholesky square root which is a triangular matrix.

The multivariate normal distribution

We say that an n-dimensional random vector $\mathbf{Y}$ has multivariate normal distribution with mean vector μ and covariance matrix Σ and write

$$\mathbf{Y} \sim N_n(\mu, \mathbf{\Sigma})$$

if $\mathbf{Y}$ has joint probability density function (pdf)

$$f(\mathbf{y}) = (2\pi)^{-n/2} |\mathbf{\Sigma}|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{y}-\mu)' \mathbf{\Sigma}^{-1}(\mathbf{y}-\mu)\}, \quad \forall \mathbf{y}$$

Note that this function has the two worst things in matrices, the determinant and the inverse of a matrix.

For this reason people often prefer to characterize the normal distribution by the moment generating function (mgf)

$$M(\mathbf{t}) = Ee^{\mathbf{Y}'\mathbf{t}} = \exp\left(\mu'\mathbf{t} + \frac{1}{2}\mathbf{t}'\mathbf{\Sigma}\mathbf{t}\right)$$

Note that the mgf is essentially the Laplace transform of the density function which is $M(-\mathbf{t})$. If we were going to derive properties of multivariate normal distribution, we would use the mgf.

Thre important properties of multivariate normal are the following

1. (basic fact about multivariate normal) $\mathbf{Y} \sim N_n(\mu, \mathbf{\Sigma})$ implies

$$\mathbf{AY} + \mathbf{b} \sim N_q(\mathbf{A}\mu + \mathbf{b}, \mathbf{A\Sigma A}')$$

2. $\mathbf{Y} \sim N_n(\mu, \mathbf{\Sigma})$ implies $Y_i \sim N_1(\mu_i, \Sigma_{ii})$.

3. If U and V are jointly normally distributed, then $\text{cov}(U,V) = 0 \Rightarrow U$ and V are independent.

We now give a brief digression on how to simulate a multivariate normal. Let $Z_{1,} \ldots Z_n$ be independent random variables, $Z_i \sim N(0,1)$,

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} \sim N_n(0, \mathbf{I})$$

Let

$$\mathbf{Y} = \mathbf{\Sigma}^{1/2}\mathbf{Z} + \mu \sim N(\mu, \mathbf{\Sigma})$$

by the basic result above when $\mathbf{\Sigma}^{1/2}$ is a square root of the non-nonnegative definite matrix $\mathbf{\Sigma}$. This shows that a multivariate normal distribution exists for any μ and any non-negative definite matrix $\mathbf{\Sigma}$, and gives a pretty easy was to simulate it.

Multiple linear regression

The basic model

Let $y = f(\mathbf{x})$ be a univariate function of several variables. When I was in high school (in the 50's) we called the $x's$ independent variables and the y the dependent variable. Now we call the $x's$ the *predictors* and the y the *response.* In simple linear regression we have one predictor and one response; in multiple linear regression we have several predictors and one response; and in multivariate linear regression we have several predictors and several responses. In this tutorial we will look at multiple linear regression with simple linear regression as a special case.

We assume that we have some data. Let Y_i be the response for the ith data point and let $\mathbf{x}_i$ be the p-dimensional (row vector) of the predictors for the ith data point, $i = 1, \cdots n$.

We assume that

$$Y_i = \mathbf{x}_i\beta + e_i.$$

Note that β is $p \times 1$. and is an unknown parameter.

For the regression model we assume that

$$e_i \sim N_1\left(0, \sigma^2\right), \text{ and the } e_i \text{ are indpendent.}$$

Note that σ^2 is another parameter for this model.

We further assume that the predictors are linearly independent. Thus we could have the second predictor be the square of the first predictor, the third one the cube of the first one, etc, so this model includes polynomial regression.

We often write this model in matrices. Let

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}, \ \mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

so that $\mathbf{Y}$ and $\mathbf{e}$ are $n \times 1$ and $\mathbf{X}$ is $n \times p$. The assumed linear independence of the predictors implies that the columns of $\mathbf{X}$ are linearly independent and hence rank$(\mathbf{X}) = p$. The normal model can be stated more compactly as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}, \ \ \mathbf{e} \sim N_n \left(0, \sigma^2 \mathbf{I}\right)$$

or as

$$\mathbf{Y} \sim N_n \left(\mathbf{X}\beta, \sigma^2 \mathbf{I}\right)$$

Therefore, using the formula for the multivariate normal density function, we see that the joint density if the observations is

$$\begin{aligned} f_{\beta,\sigma^2}(\mathbf{y}) &= (2\pi)^{-n/2} \left|\sigma^2 \mathbf{I}\right|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)' \left(\sigma^2 \mathbf{I}\right)^{-1} (\mathbf{y} - \mathbf{X}\beta)\} \\ &= (2\pi)^{-n/2} \left(\sigma^2\right)^{-n/2} \exp\{-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\beta\|^2\} \end{aligned}$$

Therefore the likelihood for this model is

$$L_{\mathbf{Y}}\left(\beta,\sigma^2\right) = (2\pi)^{-n/2} \left(\sigma^2\right)^{-n/2} \exp\{-\frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{X}\beta\|^2\}$$

Estimation of β

We first mention than the assumption on the $\mathbf{X}$ matrix implies that $\mathbf{X}'\mathbf{X}$ is invertible.

The ordinary least square (OLS) estimator of β is found by minimizing

$$q(\beta) = \sum (Y_i - \mathbf{x}_i\beta)^2 = \|\mathbf{Y} - \mathbf{X}\beta\|^2$$

The formula for the OLS estimator of β is

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

To see this note that

$$\nabla q(\beta) = 2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta) = 2(\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\beta)$$

setting this equal to 0 we get the above formula for $\widehat{\beta}$. For an algebraic derivation note that

$$\begin{aligned} q(\beta) &= \left\|\mathbf{Y} - \mathbf{X}\widehat{\beta}\right\|^2 + \left\|\mathbf{X}\widehat{\beta} - \mathbf{X}\beta\right\|^2 \\ &\geq \left\|\mathbf{Y} - \mathbf{X}\widehat{\beta}\right\|^2 = q\left(\widehat{\beta}\right) \end{aligned}$$

Although this is the formula we shall use for the OLS estimator, it is not how it is computed by most software package which solve the normal equations

$$\mathbf{X}'\mathbf{X}\widehat{\beta} = \mathbf{X}'\mathbf{Y}$$

typically using the sweep algorithm.

Note that

$$\begin{aligned} E\widehat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\mathbf{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \beta \\ cov\left(\widehat{\beta}\right) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\mathbf{M} \end{aligned}$$

Therefore

$$\widehat{\beta} \sim N_p\left(\beta, \sigma^2\mathbf{M}\right)$$

Therefore we note that the OLS, $\widehat{\beta}$, is an unbisaed estimator of β ($E\widehat{\beta} = \beta$) and the

$$var\left(\widehat{\beta}_i\right) = \sigma^2 M_{ii}$$

We now give some further properites of the OLS estimator.

1. (Gauss-Markov) For the non-normal model the OLS estimator is the best linear unbised estimator (BLUE), i.e., it has smaller variance than any othe linear unbiased estimator.

2. For the normal model, the OLS is the best unbiased estimator i.e., has smaller variance than any other unbiased estimator

3. Typically, the OLS estimator is consistent, i.e. $\widehat{\beta} \to \beta$

The unbiased estimator of σ^2

In regression we typically estimate σ^2 by

$$\widehat{\sigma}^2 = \left\| \mathbf{Y} - \mathbf{X}\widehat{\beta} \right\|^2 / (n-p)$$

which is called the unbiased estimator of σ^2. we first state the distribution of $\widehat{\sigma}^2$.

$$\frac{(n-p)\,\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p} \text{ independently of } \widehat{\beta}$$

We now give some properties of this estimator

1. For the general model $\widehat{\sigma}^2$ is unbiased

2. For the normal model $\widehat{\sigma}^2$ is the best unbiased estimator.

3. $\widehat{\sigma}^2$ is consistent

The maximum likelihood estimator (MLE)

Looking at the likelihood above, we see that the OLS estimator maximizes the exponent so that $\widehat{\beta}$ is the MLE of β. To find the MLE of σ^2 differentiate $\log\Big(L_Y\left(\widehat{\beta}, \sigma^2\right)\Big)$ with respect to σ, getting

$$\widehat{\sigma}^2_{MLE} = \frac{n-p}{n}\widehat{\sigma}^2$$

Note that if

$$p/n = q$$

then

$$E\widehat{\sigma}^2_{MLE} = (1-q)\,\sigma^2,\ \widehat{\sigma}^2_{MLE} \to (1-q)\,\sigma^2$$

so the MLE is not unbiased and is not consistent unless $p/n \to 0$.

Interval estimators and tests.

We first discuss infererence about β_i the ith component of β. Note that $\widehat{\beta}_i$ the ith component of the OLS estimator is the estimator of β_i. Further

$$var\left(\widehat{\beta}_i\right) = \sigma^2 M_{ii}$$

which implies that that the standard error of $\widehat{\beta}_i$ is

$$\widehat{\sigma}_{\widehat{\beta}_i} = \widehat{\sigma}\sqrt{M_{ii}}$$

Therefore we see that a $1-\alpha$ confidence interval for β_i is

$$\beta_i \in \widehat{\beta}_i \pm t^{\alpha/2}_{n-p}\widehat{\sigma}_{\widehat{\beta}_i}.$$

To test the null hypothesis $\beta_i = c$ against one and two-sided alternatives we use the t-statistic

$$t = \frac{\widehat{\beta}_i - c}{\widehat{\sigma}_{\widehat{\beta}_i}} \sim t_{n-p}.$$

Now consider inference for $\delta = \mathbf{a}'\beta$, let

$$\widehat{\delta} = \mathbf{a}'\widehat{\beta} \sim N_1\left(\delta, \sigma^2\mathbf{a}'\mathbf{Ma}\right)$$

therefore we see that $\widehat{\delta}$ is the estimator of δ, and

$$var\left(\widehat{\delta}\right) = \sigma^2\mathbf{a}'\mathbf{Ma}$$

so that the standard error of $\widehat{\delta}$ is

$$\widehat{\sigma}_{\widehat{\delta}} = \widehat{\sigma}\sqrt{\mathbf{a}'\mathbf{Ma}}$$

and therefore the confidence interval for δ is

$$\delta \in \widehat{\delta} \pm t^{\alpha/2}_{n-p}\widehat{\sigma}_{\widehat{\delta}}$$

and the test statistic for testing $\delta = c$ is given by

$$\frac{\widehat{\delta} - c}{\widehat{\sigma}_{\widehat{\delta}}} \sim t_{n-p} \text{ under the null hypothesis}$$

There are tests and confidence regions for vector generalizations of these procedures.

Let $\mathbf{x}_0$ be a row vector of predictors for an new response Y_0. Let $\mu_0 = \mathbf{x}_0\beta = EY_0$. The $\widehat{\mu}_0 = \mathbf{x}_0\widehat{\beta}$ is the obvious estimator of μ_0 and

$$var\left(\widehat{\mu}_0\right) = \sigma^2 \mathbf{x}_0 \mathbf{M} \mathbf{x}_0' \Rightarrow \widehat{\sigma}_{\widehat{\mu}_0} = \widehat{\sigma}\sqrt{\mathbf{x}_0 \mathbf{M} \mathbf{x}_0'}$$

and therefore a confidence interval for μ_0 is

$$\mu_0 \in \widehat{\mu}_0 \pm t_{n-p}^{\alpha/2} \widehat{\sigma}_{\widehat{\mu}_0}$$

A $1-\alpha$ prediction interval for Y_0 is an interval such that

$$P\left(a\left(\mathbf{Y}\right) \leq Y_0 \leq b\left(\mathbf{Y}\right)\right) = 1 - \alpha$$

A $1-\alpha$ prediction interval for Y_0 is

$$Y_0 \in \widehat{\mu}_0 \pm t_{n-p}^{\alpha/2} \sqrt{\widehat{\sigma}^2 + \widehat{\sigma}_{\widehat{\mu}_0}^2}$$

The derivation of this interval is based on the fact that

$$var\left(Y_0 - \widehat{\mu}_0\right) = \sigma^2 + \sigma_{\widehat{\mu}_0}^2$$

The hat matrix

The hat matrix $\mathbf{H}$ is defined as

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}$$

$\mathbf{H}$ is a symmetric idempotent matrix, i.e

$$\mathbf{H}' = \mathbf{H},\ \mathbf{H}^2 = \mathbf{H}$$

Let $\mu = \mathbf{X}\beta$, $\widehat{\mu} = \mathbf{X}\widehat{\beta}$. Then

$$\widehat{\mu} = \mathbf{HY}$$

which is why $\mathbf{H}$ is called the hat matrix. Now let

$$\mathbf{H}^{\perp} = \mathbf{I} - \mathbf{H}$$

then $\mathbf{H}^{\perp}$ is also a symmetric idempotent matrix which is orthogonal to $\mathbf{H}$, i.e

$$\mathbf{H}'\mathbf{H}^{\perp} = \mathbf{0}$$

Then

$$(n-p)\,\widehat{\sigma}^2 = \left\|\mathbf{H}^{\perp}\mathbf{Y}\right\|$$

Note that

$$\mathbf{Y} = \mathbf{HY} + \mathbf{H}^{\perp}\mathbf{Y}$$

We think of think of $\mathbf{HY}$ as having information about the signal μ and $\mathbf{H}^{\perp}\mathbf{Y}$ as having information about the noise $Y - \mu$. For the rest of this talk, we shall use $\mathbf{H}$ for this matrices

$\mathbf{R}^2$, adjusted $\mathbf{R}^2$ and predictive $\mathbf{R}^2$

Let

$$T^2 = \sum \left(Y_i - \overline{Y}\right)^2,\ S^2 = \left\|\mathbf{Y} - \mathbf{X}\widehat{\beta}\right\|^2$$

be the numerators of the variance estimators for the regression model and the intercept only model. We think of these as measuring the "variation" under these two models. Then the coefficient of determination R^2 is defined by

$$R^2 = \frac{T^2 - S^2}{T^2}$$

Note that

$$0 \le R^2 \le 1$$

Note that $T^2 - S^2$ is the amount of variation in the intercept only model which has been explained by including the extra predictors of the regression model and R^2 is the proportion of the variation left in the intercept only model which has been explained by including the additional predictors.

Note that

$$R^2 = \frac{\frac{T^2}{n} - \frac{S^2}{n}}{\frac{T^2}{n}}$$

which suggests that this might be improved by substituting unbiased estimator for the MLE's getting adjusted R^2

$$R_a^2 = \frac{\frac{T^2}{n-1} - \frac{S^2}{n-p}}{\frac{T^2}{n-1}} = 1 - \frac{n-1}{n-p}\left(I - R^2\right)$$

Both R^2 and adjusted R^2 suffer from the fact that the fit is being evaluated with the same data used to compute it and the therefore the fit looks better than it is. A better procedure is based on cross-validation. Suppose we delete the ith observation and compute $\widehat{\beta}_{-i}$ the OLS estimator of β without the ith observation. We do this for all i. We also compute $\overline{Y}_{-i}$

$$\overline{Y}_{-i} = \sum_{j \neq i} Y_j / \left(n-1\right)$$

the sample mean of theY_i without the ith one. Then let

$$T_p^2 = \sum \left(Y_i - \overline{Y}_{-i}\right)^2 = \frac{nT^2}{n-1}$$

$$S_p^2 = \sum \left(Y_i - \mathbf{x}_i \widehat{\beta}_{-i}\right)^2 = \sum \left(\frac{Y_i - \mathbf{x}_i \widehat{\beta}}{1 - H_{ii}}\right)^2$$

(where H_{ii} is the ith diagonal of the hat matrix).

Then predictive R^2 is defined as

$$R_p^2 = \frac{T_p^2 - S_p^2}{T_p^2}$$

Predictive R^2 computes the fit to the ith observation without using that observation and is therefore a better measure of the fit of the model that R^2 or adjusted R^2.

Diagnostics

Residuals

Most of the assumptions in regression follow from

$$\mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \mathbf{Y} - \mathbf{X}\beta \sim N_n\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$$

To check these assumptions we look at residuals The ordinary residuals are

$$\widehat{\mathbf{e}} = \begin{pmatrix} \widehat{e}_1 \\ \vdots \\ \widehat{e}_n \end{pmatrix} = \mathbf{Y} - \mathbf{X}\widehat{\beta} =$$

$$\left(\mathbf{I} - \mathbf{H}\right)\mathbf{Y} \sim N_n\left(0, \sigma^2\left(\mathbf{I} - \mathbf{H}\right)\right)$$

Note that the e_i are assumed to have equal variances, but even if all the assumptions are met

$$var\left(\widehat{e}_i\right) = \sigma^2\left(1 - H_{ii}\right)$$

are different. For this reason, the residual are often standardized getting the standardized residuals

$$\widehat{e}_{is} = \frac{\widehat{e}_i}{\widehat{\sigma}\sqrt{1 - H_{ii}}}$$

which if the assumptions are met should have constant variance about 1.

Because of the unequal variances, the ordinary residuals can be misleading, so we always look at the standardized residuals. Many other type of residuals have been suggested, e.g. delete one residual and t residual but they seem to look just like standardized residuals and so it does not seem necessary to look at any other residuals but standardized residuals. Just don't use ordinary residuals.

The assumption on the errors is really 4 assumptions

1. $Ee_i = 0$. This means we have included enough terms in the model. If it is not satisfied, it can often be corrected by including more terms in the model. This is often

a tough assumption to check with residuals, since it can be shown that the average of the residuals is always 0, even if this assumption is violated. One situation where residuals can be useful is in polynomial regression on a variable x. In that case if we plot the residuals against x, and if we have too few terms, we should see a pattern.

2. $var\,(e_i)$ is constant. This is the most important assumption and is often violated. One way to use residuals to check this assumption is to make a residual vs. fits plot. For example, if we see a fanning pattern with large residuals vs. large fits, this means the variance is increasing with the mean. If we see this it is often remedied by a log transformation on the Y_i. Another way to go is to use weighted least squares.

3. The e_i are independent. This is another important assumption which is hard to check with residuals. If it is not true, we can model the correlation between the observations using time series methods or repeated measures or generalized least squares.

4. The e_i are normally distributed. This is the least important assumption. For moderate sample sizes it has been shown that that regression is robust against the normal assumption.To use residuals to check this assumption, look at a normal scores plot of the (standardized) residuals. It should look like a straight line. If this assumption is not met, you can transform to achieve normality, you can use an M-estimator, an R-estimator or some other less sensitive estimator than OLS or you can ignore it.

One other use for residual is for looking for outliers, points whose observations seem incorrect. One rule is that an observation is an outlier if its absolute standardized residual is greater than 3. Some data analysis programs automatically eliminate all outliers from the data.

One (true) story that suggest that this is not a good idea has to do with the hole in the ozone, which was not discovered by satelite (as it should have been), because the data analysis programs used eliminated all outliers and so eliminated the data for the hole in the ozone. It was discovered from the ground much later than it would have been discovered by satelite if the data had not been cleaned.

I always say you should look carefully at the outliers and think about them before you eliminate them. Several times in my consulting the outliers were the most interesting points in the data. We often do separate analyses on the outliers and learn things we coul not learn from the clean data. Basically, before you elimate an outlier, you try to decide if it is a mistake or an unusual data point. If it is a mistake, eliminate it, if it is an unusual data point then try to learn from it.

Influence

Often the values for the predictors for one observation are quite far from the other observations which leads to that observation having a large influence on the regression line. For example in a simple regression, we might have most of the observatiions with predictor about 10 and one observation with predictor 10^{10}. Then the regression line will basically connect the one extreme observation wiht the middle of the cloud of other points, so the response assocaited withe extreme point will essentially determine the regression line.

The leverage of the ith observation is defined as H_{ii}, the ith diagonal of the hat matrix. The reson for this definition is that if $\mu = \mathbf{X}\beta$,then

$$\widehat{\mu} = \mathbf{HY}$$

so that the ith diagonal element of the hat matrix is the coefficient of the ith observation in its estimated mean. If this coefficient is large, then the ith observation has a large influence on its estimated mean and if the coefficient is small, then the ith observation has little influence on its estimated mean.

Using the fact that $\mathbf{H}$ and $\mathbf{I} - \mathbf{H}$ are idempotent and hence non-negative definite, we can show that

$$0 \leq H_{ii} \leq 1$$

so an observation is influential if the influence near 1 and not if it near 0. Note also that

$$\sum h_{ii} = tr\mathbf{H} = tr\left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}\right) = tr\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\right) = tr\mathbf{I}_p = p$$

so that the average leverage is

$$\overline{H} = \sum H_{ii}/n = p/n$$

One rule of thumb which is often used is that an observation has high influence if

$$H_{ii} > \frac{3p}{n}.$$

If we find a point which has high influence, we should think about whether we should eliminate it. Sometimes such a point has an incorrect number for the predictor and could really mess up the analysis. Sometimes. however, it is a true point and may be the most important point in fitting the regression.

Multicolinearity

One other critical assumption in regession is that the predictors linearly independent so that $\mathbf{X}'\mathbf{X}$ is invertible. Typically this assumption is satisfied. Often though one predictor is nearly a linear combination of some others. This is called multicollinearity. When this happens the $var\left(\widehat{\beta}_i\right)$ are quite large and it is not possible to draw good inference about the β_i. So we try to detect multicolinearity and eliminate it.

The main tool for detecting multicolinearity is the variance inflation factor (VIF) for each predictor which we now describe. Recall that

$$var\left(\widehat{\beta}_i\right) = \sigma^2 M_{ii}$$

We say that the predictors are orthogonal if for any two columns of the $\mathbf{X}$ matrix

$$\mathbf{X}_j'\mathbf{X}_k = 0,\ \forall j = k$$

We note that orthogonality is as far from multicolinearity as possible. We note if the predictors are orthogonal then

$$var_O\left(\widehat{\beta}_i\right) = \sigma^2 / \left\|\mathbf{X}_i\right\|^2$$

The VIF for the ith predictor is defined as

$$\frac{var\left(\widehat{\beta}_i\right)}{var_O\left(\widehat{\beta}_i\right)}$$

so the VIF tells how much the variance of $\widehat{\beta}_i$ has been inflated due to the multicolinearity. If it is large then something should probably be done to eliminate the multicolinearity. If it they are all near 1 then there is no multicolinearity.

There is another interpretation for VIF's which is pretty interesting. Suppose we regressed the jth predictor on the other predictors and let R_j^2 be R^2 from this fit. Then, it can be shown that

$$VIF_j = \frac{1}{1 - R_j^2}$$

so that if the jth predictor is nearly a linear combination of the others then R_j^2 should be near 1 and the VIF_j should be large.

Typically in a polynomial regression model fit in the obvious way there is a great deal of colinearity. One method which is often used to eliminate the colinearity in this situation is to center the x term for the linear term, then square the centered x's for the quadratic term, etc.

Model Selection

The last regression topic we'll talk about is how to chose which pedictors to include in the model. We say we have overfit the model if we have too many terms and undrfit it if we have too few terms.

Some naive approaches don't work, such as choosing the model with the largest R^2. It can be shown that R^2 always increases when variables are added to the model and we end up by including all the predictors in the model which is usually extreme overfitting. Maximizing adjusted R^2 is a little better, but stilll overfits. Maximizing predictive R^2 seems to work reasonably well.

Another aspproach which is often used is to minimize Mallow's C$_p$, which we now describe. Let

$$Q = \frac{E\left\|\widehat{\mu} - \mu\right\|^2}{\sigma^2} = p + \frac{\widehat{\mu}\left(\mathbf{I} - \mathbf{H}\right)\widehat{\mu}}{\sigma^2}$$

Our goal is to find a model which minimizes Q. It can be shown that an unbiased estimator of Q is

$$\widehat{Q} = \frac{(n-p)\,\widehat{\sigma}^2}{\sigma^2} - n + 2p.$$

$\widehat{Q}$ is called Mallows C_p. We can already compute all of this except σ^2,which we estimate by regessing on all the possible predictors. Then we look at all the possible models and find the one which minimizes $\widehat{Q}$. The main problem with this appraoch is the estimation of σ^2. Sometimes there are more potential prectors than there are observations so it not possible to regress on all possible predictors. Also it seems bothersome that if we add more predictors to the model, we would change σ^2. It seems that the criterion for a particular model should depend only on that model not some larger model.

For these reasons, emphasis for model selection has shifted to penalized likelihood criteria. Note that for this model, the maximized likelhood is

$$L_Y\left(\widehat{\beta}, \widehat{\sigma}^2_{MLE}\right) = (2\pi)^{-\frac{n}{2}}\left(\widehat{\sigma}^2_{MLE}\right)^{-\frac{n}{2}} \exp\{-\frac{\left\|\mathbf{Y} - \mathbf{X}\widehat{\beta}\right\|^2}{2\widehat{\sigma}^2_{MLE}}\}$$

$$= (2\pi)^{-\frac{n}{2}}\left(\widehat{\sigma}^2_{MLE}\right)^{-\frac{n}{2}} \exp\{-\frac{n}{2}\}$$

A naive approah would be to choose the model which maximizes the maximized likehood, but that also just picks out the model with all the predictors and overfits.

The first penalized likelihood criterion suggested was the Akaike Information Criterion (AIC), which minimizes

$$AIC = -2\log\left(L_Y\left(\widehat{\beta}, \widehat{\sigma}^2_{MLE}\right)\right) + 2\left(p+1\right)$$

This criterion is based on Kullback-Liebler information. Unfortunately, it is known to overfit.

A better penalized likelihood citerion is Shwartz' Bayesian Information Criterion (BIC) which minimizes

$$BIC = -2\log\left(L_Y\left(\widehat{\beta}, \widehat{\sigma}^2_{MLE}\right)\right) + \left(p+1\right)\log n$$

This criterion is derived using Bayesian ideas and is known to be consistent, in that it finds the correct model as $n \to \infty$. BIC seems to be the method which people are using more and more although some regression software programs do not compute either AIC or BIC.

Some fallacies in regression

Association is not causation

Many people seem to believe that if a predictor is significant then the that predictor is causing the response. Due to the observational nature of regression it not possible to establish causation. A couple of examples of weird conclusions from the faculty senate here at Penn State may be helpful

1. A study of writing skills of Penn State students concluded that the best predictor of improve writing skills was whether the student had taken a basic stat course. So many

of the faculty concluded that is we wanted to improve students writing skills we should have them take a basic stat couse, a course with no writing training. This rather stupid conclusions is an example of using association to establish causation. What probably happenned is that the students who improved their writing skills the most were social scientists who are required to take a basic stat course.

2. A study of student performance at Penn State found that the best predictor of satisfactory performance at Penn State was having take an AP course in calculus in high school. It was suggested therefore if we wanted to get good students from schools in downtown Philadelphia, all we needed to do was teach them an AP course in calculus. This is again the mistake of using association to establish causation. What probably happenned is the student who had an AP course was a good student at a good high school and naturally did well at Penn State.

The assumption that association implies causation is an insidious one and very easy to fall into. I have seen many intelligent people make this mistake. Even as a statistician I have made this mistake myself.

The regression fallacy (Regression to the mean)

An early example of this phenomenon was in Galton's original work on regression where he developed many of the ideas of regression. He regressed son's height on father's height. What he found was that very tall fathers had sons who were shorter than they were but still tall and very short fathers had sons who were taller then they were but still short. He interpreted this as due to a natural mechanism to control the numbers of outliers. But this is just due to the random errors. A very tall person probably has the signal for tallness but also the noise leads him to be a little bit taller still. His son inherits the signal part but not the noise, so he is a little shorter. For example, I doubt that many people would expect Yao MIng's children to be as tall as he is.

Another example which has been studied occurs in baseball, where it is often observed that the people with extraordinary high batting averages at midseason usually have lower batting averages by the end of the season. This is because, the reason that he had a super high batting average at mid season is because he is a good hitter and because he had an usually large number of lucky hits. The lucky hits often even out be the end of the season.

Once you are attuned to this phenomenon, you start to see regession to the mean everywhere.