Simple Linear Regression

Parameters

The simple model is

$$y_i = \beta_0 + \beta_1 x_i$$

with $E(y_i) = \beta_0 + \beta_1 x_i$,$\text{Var}(y_i) = \text{Var}(\beta_0 + \beta_1 x_i + \epsilon) = \sigma^2$

Estimates for β_0, β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n k_i y_i = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$k_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^n y_i (x_i - \bar{x})$$

$$\text{Var}(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^n k_i^2$$

Estimation on σ^2

$$SSE = \sum_{i=1}^n e_i^2 = (y_i - \hat{y}_i)^2$$

SSE has $n-2$ degrees of freedom.

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = MSE$$

Hypothesis Testing on the Parameters

Testing on the slope for a constant β

$$H_0 : \hat{\beta}_1 = \beta, \;\; H_1 : \hat{\beta}_1 \neq \beta$$

If σ^2 is known,

$$Z_0 = \frac{\hat{\beta}_1 - \beta}{\sqrt{\sigma^2 / S_{xx}}}$$

If σ^2 is unknown,

$$t_0 = \frac{\hat{\beta}_1 - \beta}{\sqrt{MSE / S_{xx}}}$$

We reject the null hypothesis $|t_0| > t_{\alpha/2, n-2}$. We test the intercept similarly,

$$H_0 : \beta_0 = \beta, \;\; H_1 : \beta_0 \neq \beta$$

$$t_0 = \frac{\hat{\beta}_0 - \beta}{se(\hat{\beta}_0)}$$

where $se^2(\hat{\beta}_1) = \frac{MSE}{S_{xx}}$, $se^2(\hat{\beta}_0) = MSE(1/n + \bar{X}^2 / S_{xx})$

Significance of Regression

We test signficance with

$$H_0 : \beta_1 = 0, \;\; H_1 : \beta_1 \neq 0$$

Using the same statistic, $|t_0| > t_{\alpha/2, n-2}$.

Source	SS	DF
Regression	$SSR = \hat{\beta}_1^2 \sum (X_i - \bar{X})^2$	$p-1$
Error	$SSE = \sum (Y_i - \hat{Y}_i)^2$	$n-p$
Total	$SSTO = \sum (Y_i - \bar{Y})^2$	$n-1$

MS=SS/df	E(MS)	F
MSR	$\sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$	MSR/MSE
MSE	σ^2	

Confidence Intervals

The confidence interval on the slope β_1,

$$\hat{\beta}_1 - t_{\alpha/2, n-2} se(\hat{\beta}_1) \leq \hat{\beta}_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} se(\hat{\beta}_1)$$

For the intercept β_0,

$$\hat{\beta}_0 - t_{\alpha/2, n-2} se(\hat{\beta}_0) \leq \hat{\beta}_0 \leq \hat{\beta}_0 + t_{\alpha/2, n-2} se(\hat{\beta}_0)$$

For σ^2,

$$\frac{(n-2)MSE}{\chi^2_{\alpha/2, n-2}} \leq \sigma^2 \leq \frac{(n-2)MSE}{\chi^2_{1-\alpha/2, n-2}}$$

To

Interval Estimation on Mean Response

An unbiased estimator for $E(y|x_0)$ for a value of regressor $x = x_0$ is

$$\widehat{E(y|x_0)} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

The variance is

$$\text{Var}(\hat{\mu}_{y|x_0}) = \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{S_{xx}}$$

The sampling distribution for

$$\frac{\hat{\mu}_{y|x_0} - E(y|x_0)}{\sqrt{MSE(1/n + (x_0 - \bar{x})^2 / S_{xx})}} \sim t_{n-2}$$

So the confidence interval is then

$$\left[\hat{\mu}_{y|x_0} \pm t_{\alpha/2, n-2} \sqrt{MSE \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} \right]$$

Prediction of New Observations

If x_0 is the new value for x, then $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ is the point estimate for the response. The new error is

$$\psi = y_0 - \hat{y}_0 \implies \text{Var}(\psi) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

Then we use the standard error of ψ to construct the prediction interval

$$\left[\hat{y}_0 \pm t_{\alpha/2, n-2} \sqrt{MSE \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} \right]$$

To hypotheses $H_0 : y_0 = y_{00}$, $H_1 : y_0 \neq y_{00}$, reject null hypothesis when $|t_0| > t_{\alpha/2, n-2}$

$$\frac{y_0 - y_{00}}{\sqrt{MSE \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}} \sim t_{n-2}$$

Correlation

The coefficient of determination is

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The adjusted R^2 value is

$$R^2_{Adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

The pearson correlation coefficient is

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}}$$

When applied to a sample,

$$r = b_1 \left(\frac{S_{xx}}{SST} \right)^{\frac{1}{2}} = \frac{S_{xy}}{(S_{xx} SST)^{1/2}}$$

If we want to test $H_0 : \rho = 0$, $H_1 : \rho \neq 0$, use the t statistic,

$$t = \frac{r \sqrt{n-2}}{\sqrt{1 - r^2}} \sim t_{n-2}$$

We reject the null hypothesis $H_0 : \rho = 0$ if $|t_0| > t_{\alpha/2, n-2}$. To test $\rho = \rho_0$,

$$H_0 : \rho = \rho_0, \;\; H_1 : \rho \neq \rho_0$$

Use the standardized test statistic

$$Z_0 = (\text{arctanh}(r) - \text{arctanh}(\rho_0)) \sqrt{n-3}$$

We can obtain our confidence interval with

$$\left[\tanh \left(\text{arctanh}(r) \pm \frac{Z_{\alpha/2}}{\sqrt{n-3}} \right) \right]$$

where $\tanh(u) = (e^u - e^{-u})/(e^u + e^{-u})$. We reject $H_0 : \rho = \rho_0$ if $|Z_0| > Z_{\alpha/2}$.

Multiple Linear Regression

Model

We write the multiple linear regression model as

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

where (Note $p = k + 1$.)

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{11} & \cdots & x_{1k} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Y is $n \times 1$, X is $n \times p$, $\boldsymbol{\beta}$ is $p \times 1$, and $\boldsymbol{\epsilon}$ is $n \times 1$. In matrix form, we get the fitted line

$$\hat{Y} = X\hat{\boldsymbol{\beta}} = X(X'X)^{-1}X'Y = HY$$

$H = X(X'X)^{-1}X'$ is the **hat matrix**.

Properties of the Hat Matrix

(a) H is a projection matrix, so it is idempotent and symmetric $HH = H$, $H' = H$.

(b) The matrix H is orthogonal to the matrix $I - H$, so $(I - H)H = H - HH = 0$. Moreover, $(I - H)$ is idempotent and a project matrix as well.

(c) The vector of residuls is

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

(d) Y is projected onto a space spanned by the columns of H, and the residuals are in an orthogonal space.

$$Y = HY + (I - H)Y$$

Estimation of σ^2

Residual sum of squares is

$$SSE = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}'\boldsymbol{e} = Y'Y - \hat{\boldsymbol{\beta}}'X'Y$$

SSE has $n - p$ degrees of freedom, then MSE is

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n - p}$$

Estimation and Hypothesis Testing

Testing for Significance

We test for significance with

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0, \; H_1 : \beta_j \neq 0$$

Rejecting the null hypothesis means at least one regressor contributed signficantly. We use an F statistc

$$F_0 = \frac{SSR/k}{SSE/(n-p)} = \frac{MSR}{MSE} \sim F_{k,n-p}$$

We reject the null hypothesis when $F_0 > F_{\alpha,k,n-k-1}$.

The **total sum of squares** is

$$SST = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2 = Y'Y - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)$$

The **regression sum of squares** is

$$SSR = \hat{\beta}'X'Y - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)$$

The **residual sum of squares** is

$$SSE = Y'Y - \hat{\beta}'X'Y = Y'(I - H)Y$$

We can also write SST and SSR in terms of the J_n, and $n \times n$ matrix with 1's.

$$SST = Y'\left(I - \frac{1}{n}J_n\right)Y$$

$$SSR = Y'\left(H - \frac{1}{n}J_n\right)Y$$

Tests on Individual Coefficients

To test an indivual coefficient β_j, we use

$$H_0 : \beta_j = 0, \; H_1 : \beta_j \neq 0$$

The test statistic is

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

Where C_{jj} is the diagonal entry of $(X'X)^{-1}$. We reject H_0 when $|t_0| > t_{\alpha/2,n-p}$.

If we fail to reject the null hypothesis, we can remove the corresponding regressor x_j from the model.

Extra Sum Of Squares

We want to partition r of the k regressors to test

$$H_0 : \beta_2 = 0, H_1 : \beta_2 \neq 0$$

$Y = X\beta + \epsilon$,where Y is $n \times 1$, X is $n \times p$, β is $p \times 1$, and ϵ is $n \times 1$ with $p = k + 1$.

Full Model

$$Y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon$$

X_1 is $n \times (p - r)$, X_2 is $n \times r$.

$$\hat{\beta} = (X'X)^{-1}X'Y, \; SSR(\beta) = \hat{\beta}'X'Y$$

which has $k = p - 1$ degrees of freedom, $df_F = n - p$.

Reduced Model

To test regressors in β_2, fit the model assuming $H_0 : \beta_2 = 0$ is true.

$$Y = X_1\beta_1 + \epsilon$$

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y, \; SSR(\beta_1) = \hat{\beta}_1{}'X_1'Y$$

which has $k - r = p - 1 - r$, $df_R = n - p + r$ degrees of freedom. The sum of squares due to β_2 given that β_1 is already in the model is

$$SSR(\beta_2|\beta_1) = SSR(\beta) - SSR(\beta_1)$$

The null hypothesis $\beta_2 = 0$ can be tested with (partial F-test)

$$F_0 = \frac{SSR(\beta_2|\beta_1)/r}{MSE} = \frac{\frac{SSE(R)-SSE(F)}{df_R - df_F}}{MSE}$$

If $F_0 > F_{\alpha,r,n-p}$, then we reject the null hypothesis and conclude that at least one regressor in X_2 contributes.

Lack of Fit

Pure Error Sum of Squares:

$$SS_{PE} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

Sum of Squares Due to Lack of Fit:

$$SS_{LOF} = \sum_{i=1}^{m} n_i(\bar{y} - \hat{y}_i)$$

F-Statistic:

$$F^* = \frac{SS_{LOF}/(m-2)}{SS_{PE}(n-m)} = \frac{MS_{LOF}}{MS_{PE}}$$

Testing Lack of Fit

If the regression is linear, then $E(y_i) = \beta_0 + \beta_1 x_i$,

$$H_0 : E(y_i) = \beta_0 + \beta_1 x_i, \; H_1 : E(y_i) \neq \beta_0 + \beta_1 x_i$$

Reject the null hypothesis when $F^* > F_{\alpha,m-2,n-m}$.

Anova Table for Lack of Fit

Source	Sum of Squares	DF
Regression	$SSR = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_i)^2$	1
Residuals	$SSE(R) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_i)^2$	$n-2$
Lack of Fit	$SS_{LOF} = \sum_{i=1}^{m} n_i(\bar{y}_i-\hat{y}_i)^2$	$m-2$
Pure Error	$SS_{PE} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$	$n-m$
Total	$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2$	$n-1$

Source	Mean Square = SS/df	F-Statistic
Regression	$SSR/1$	MSR/MSE
Residuals	$SSE(R)/n-2$	
Lack of Fit	$SS_{LOF}/m-2$	MS_{LOF}/MS_{PE}
Pure Error	$SS_{PE}/n-m$	

Model Adequacy

Normaility

- **Using a boxplot:** Box plot of residuals should be symmetric around a median of 0.
- **Histogram:** Should be of the shape of a normal distribution.
- **QQ-Plot:** Plot $E_k = \sqrt{MSE}\cdot\Phi^{-1}\left(\frac{k-0.375}{n+0.25}\right)$ vs the residuals $e_{(k)}$, should be a straight line.

Constant Variance

Studentize the residuals, and plot $\sqrt{e_i^*}$ vs $\hat{Y}_i$.

$$e_i^* = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

- Plot should show a random distribution of points. Otherwise, signs of non-constant variance.
- Residuals lie in a narrow band around 0 $\implies$ no need of correction.
- Residuals are increasing or decreasing $\implies$ variance is non constant.
- Double-bow pattern $\implies$ variance in the middle is larger than the variance at the extremes.
- Quadratic relationship (parabola shape) $\implies$ maybe a nonlinear relationship

Confidence Intervals

Confidence Intervals on Regression Coefficients

To construct a confidence interval on β_j, use the statistic

$$\frac{\hat{\beta}_j-\beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t_{n-p}$$

The CI is then

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_j \le \hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

Recall C_{jj} is the jth diagonal entry of $(X'X)^{-1}$ the standard error is

$$se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$$

Confidence Interval on Mean Response

To construct confidence intevrals at points $x_{01}, x_{02}, \ldots, x_{0k}$, define

$$x_0 = \begin{bmatrix}1 & x_{01} & x_{02} & \cdots & x_{0k}\end{bmatrix}^T$$

The fitting value is then

$$\hat{y}_0 = x_0'\hat{\beta}$$

This is an unbiased estimator, $E(y|x_0) = x_0'\beta = E(\hat{y}_0)$, and $\mathrm{Var}(\hat{y}_0) = \sigma^2 x_0'(X'X)^{-1}x_0$. The CI is then

$$\left[\hat{y}_0 \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1+x_0'(X'X)^{-1}x_0)}\right]$$

Simultaneous Confidence Interval

Theorem (Bonferroni Inequality). *For two events A_1, A_2, we have that*

$$P(A_1\cup A_2) \le P(A_1)+P(A_2)$$

From DeMorgan's identity, we also have

$$P(A_1^c\cap A_2^c) = 1-P(A_1\cup A_2) \ge 1-P(A_1)-P(A_2)$$

If we define the events

$$A_1^c : \hat{\beta}_0 \pm t_{\alpha/2,n-2}s(\hat{\beta}_0)$$

$$A_2^c : \hat{\beta}_1 \pm t_{\alpha/2,n-2}s(\hat{\beta}_1)$$

From Bonforroni's Inequality, if we have $P(A_1) = P(A_2) = \alpha$, then

$$P(A_1^c\cap A_2^c) \ge 1-P(A_1)-P(A_2) = 1-2\alpha$$

In general, if we have p parameters and each confidence interval has confidence, $1-\frac{\alpha}{p}$, then

$$P\left(\bigcap_{i=1}^{p} A_i^c\right) \ge 1-p\frac{\alpha}{p} = 1-\alpha$$

Transformations and Weighting

Variance Stabilizing Transformations

- **Poisson** ($\mu=\sigma^2$): $y\sim\mathrm{Poisson}(\lambda) \implies \sqrt{y}$ is nearly normal and has variance $1/4$ if λ is large.
- **Binomial:** $y\sim\mathrm{Bin}(n,p)$ with mean $m=np$, then

  $$y' = \sin^{-1}\left(\sqrt{\frac{y+c}{n+2c}}\right)$$

  The optimal value of c is $3/8$ when m and $n-m$ are large. The variance is approximately $\frac{1}{4}\left(n+\frac{1}{2}\right)^{-1}$.

Transformations to Linearize Models.

- **Exponential:** $\beta_0' = \ln\beta_0$, $\epsilon' = \ln\epsilon$,

  $$y = \beta_0 e^{\beta_1 x}\epsilon \to y' = \ln y = \beta_0' + \beta_1 x + \epsilon'$$

- **Reciprocal:** $x' = \frac{1}{x}$,

  $$y = \beta_0 + \beta_1\frac{1}{x} + \epsilon \to y = \beta_0 + \beta_1 x' + \epsilon$$

  $$\frac{1}{y} = \beta_0 + \beta_1 x + \epsilon \to y' = \frac{1}{y}$$

- **Two Step Reciprocal:** $y' = \frac{1}{y}$, $x' = \frac{1}{x}$,

  $$y = \frac{x}{\beta_0+\beta_1 x} \to y' = \beta_0 x' + \beta_1$$

Box-Cox Transformations

When data is not normally distrubted, can apply a power transformation

$$y^{(\lambda)} = \begin{cases}\frac{y^\lambda-1}{\lambda\dot{y}^{\lambda-1}} & \lambda\neq 0\\ \dot{y}\ln y & \lambda = 0\end{cases},\quad \dot{y} = \ln^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\ln y_i\right)$$

We want a value for λ that mimizes SSE, this value is found by trial and error.

Weighted Least Squares

$$W = \begin{bmatrix} w_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & w_n\end{bmatrix}$$

$$X_W = \begin{bmatrix} 1\sqrt{w_1} & \cdots & x_{1k}\sqrt{w_1}\\ 1\sqrt{w_2} & \cdots & x_{2k}\sqrt{w_2}\\ \vdots & \ddots & \vdots\\ 1\sqrt{w_n} & \cdots & x_{nk}\sqrt{w_n}\end{bmatrix},\quad Y_W = \begin{bmatrix} y_1\sqrt{w_1}\\ y_2\sqrt{w_2}\\ \vdots\\ y_n\sqrt{w_n}\end{bmatrix}$$

New Weighted Model: $Y_w = X_w\boldsymbol{\beta} + \epsilon_W$, estimate becomes

$$\hat{\boldsymbol{\beta}} = (X_W'X_W)^{-1}X_W'Y_W = (X'WX)^{-1}X'WY$$

Weighted mean square error is

$$MSE_W = \frac{\sum_{i=1}^{n} w_i(y_i-\hat{y}_i)^2}{n-p} = \frac{\sum_{i=1}^{n} w_i e_i^2}{n-p}$$

Diagnostics for Leverege

Leverge of the ith observation is defined as h_{ii},

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

We can also use the mean with the ith observation removed, $\bar{X}_{(i)}$,

$$h_{ii} = \frac{1}{n} + \left(\frac{n-1}{n}\right)^2 \frac{(X_i - \bar{X}_{(i)})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

If $h_{ii} > 2p/n$, ith observation is considered influential.

Measures of Influence

Difference in fit

Difference in fit is defined as

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = t_i \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

where t_i is the Studentized deleted residual,

$$t_i = e_i \left(\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}\right)^{\frac{1}{2}}$$

DFFITS represents the number of estimated standard deviations of $\hat{Y}_i$ that the fitted value increases or decreases. If X_i is an outlier with high liverage, then $|DFFITS_i|$ will be large. We class influential cases if

$$DFFITS_i > \begin{cases} 1 & \text{for small data sets} \\ 2\sqrt{p/n} & \text{for large data sets} \end{cases}$$

Cook's Distance

Cook's distance considers the influence of the ith observation on the entire regression line,

$$D_i = \frac{\sum_{j=1}^n (\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE} = \frac{e_i^2}{pMSE}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right)$$

D_i is large if the residual is large and leverage is moderate, or if residual is moderate and leverage is large, or both. Influential cases are $D_i > 1$.

Difference in Coefficients

DFBETAS are the differences in the estimated regression coefficients with and without the ith observation,

$$DFBETAS_{(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE_{(i)} c_{ii}}}$$

c_{ii} is the ith diagonal entry of $(X'X)^{-1}$. Large value of DFBETAS means large impact of the ith case on the kth coefficient.

$$DFBETAS_{(i)} > \begin{cases} 2/\sqrt{n} & \text{for large } n \\ 1 & \text{for small } n \end{cases}$$

Polynomial Regression

A k-order polynomial regression in one variable

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \epsilon$$

k should be as low as possible, inversion of $X'X$ will be inaccurate. Orthogonal polynomials are used to simply the fitting process,

$$Y_i = \beta_0 P_0(X_i) + \beta_1 P_1(X_i) + \beta_2 P_2(X_i) + \cdots + \beta_k X^k + \epsilon$$

where P_j is a j order polynomial

$$\sum_{i=1}^n P_j(X_i) P_l(X_i) = 0, \; j \neq l$$

$$P_0(X_i) = 1$$

Least squares estimates are given by

$$\hat{\beta}_j = \frac{\sum_{i=1}^n P_j(X_i) Y_i}{\sum_{i=1}^n P_j^2(X_i)}, \; j = 0, 1, \ldots, k$$

Advantage of this is that the model can be fitted sequentially, can be done my computers so this is not as important. With multiple variables, include them cross terms,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \epsilon$$

Indicator Regression

With qualitative, indicator functions can be used. Example of this, if you want to fit a model as a function of gender,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

With X_2 being the gender variable, so

$$Y = \begin{cases} \beta_0 + \beta_1 X_1 + \beta_2 + \epsilon & \text{Male} \\ \beta_0 + \beta_1 X_1 + \epsilon & \text{Female} \end{cases}$$

Multicolinearity

Symptoms of multicolinearity:

1. Large variation in coefficients when a new variable is added /deleted.
2. Non-significant results in individual tests on the coefficients of important variables.
3. Large coefficients of simple correlation between pairs of variables.
4. Wide confidence interval for the regression coefficients of important variables.

Variance inflation factor (VIF) is defined as

$$\text{VIF}_j = C_{jj} = (1 - R_j^2)^{-1}$$

where R_j^2 is the coefficient of multiple determinination. If $\text{VIF}_j > 10$, this is an indication that multicolinearity exists.

Detecting Multicollinearity

Consider 2 predictors X_1,X_2, if they are standardized then

$$(X'X) = \frac{1}{1 - r_{12}^2}\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$$

where r_{12} is the correlation. The covariance matrix of the coefficients is

$$\sigma^2 (X'X)^{-1} = \sigma^2 \frac{1}{1 - r_{12}^2}\begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix}$$

As $|r_{12}| \to 1$, the variance $\text{Var}(\hat{\beta}_k) \to \infty$, and the covariance $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \to \pm\infty$. The estimates are

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

which can be written as the individual estimates

$$\hat{\beta}_1 = \frac{r_{1Y} - r_{12}}{1 - r_{12}^2}, \; \hat{\beta}_2 = \frac{r_{2Y} - r_{12}}{1 - r_{12}^2}$$

Diagonal elements of $(X'X)^{-1}$ are $C_{jj} = \frac{1}{1-R_j^2}$ where R_j^2 is the R-square value obtained from the regression of X_j on the other $p-1$ variables. If there is a strong multicollinearity between X_j and the other $p-1$ variables, then

$$R_j^2 \approx 1, \; \text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{1 - R_j^2} \to \infty$$

Multicollinearity can also be detected with the mean variance inflation factor,

$$\overline{\text{VIF}} = \frac{\sum_{k=1}^{p-1} \text{VIF}_k}{p-1}$$

A value greater than 1 indicates serious multicolinearity.

Ridge Regression

A remedy for multicolinearity. Standardize normal equations to get $r_{XX}\hat{\beta} = r_{YX}$. Ridge estimator becomes $\hat{\beta}_R = (r_{XX} + cI)^{-1} r_{YX}$ for some $c \geq 0$. Using penalized least square,

$$Q = \sum (Y_i - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1})^2 + c \sum_{j=1}^{p-1} \beta_j^2$$

Mallow's C_p and Akaine Info. Criterion

Mallow's C_p statistic is given as

$$C_p = \frac{SSE_p}{MSE} - n + 2p$$

where $SSE_p = e_p' e_p$, $e_p = (1 - H_p)Y$ where H_p is the hat matrix for the p predictors. AIC is based on maximizing expected entropy, and is given as

$$\text{AIC}_p = n \ln(SSE_p) - n \ln n + 2p$$

As more variables are included, AICp decreases and the issue becomes whether or not the decrease justifies the inclusion of more variables.

Shwartz's Bayesian Criterion and PRESS

There are several Bayesian extension of AIC, such as the Shwartz's Bayesian criterion,

$$\mathrm{BIC}_{\mathrm{Sch}} = n\ln(SEE_p) - n\ln n + p\ln n$$

This criterion places a larger penalty on adding regressors as the sample size increases and is the one used in R. We can also minimize prediction sum of squares,

$$\mathrm{PRESS}_p = \sum_{i=1}^{n}(Y_i - \hat{Y}_{(i)})^2 = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$$

Techniques for Variable Selection

Forward Selection

- **Step 1:** Begin with no regressors in the model. Compute the t-statistic for each regressor and choose the greatest absolute value. A pre selection critical value F_{IN} is chosen.
- **Step 2:** Choose the next variable using the same criteria. Compute residuals from the regressions of the other regressors on X_j, that is the residuals from $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$, and $\hat{X}_j = \hat{\alpha}_{0j} + \hat{\alpha}_{1j}X_1$ for $j = 2, \ldots, K$.
- If X_2 is selected, then the largest partial F statistic is F_0 (use partial F-test), add X_2 to the model. Check to drop a variable if the t-value drops below a preset limit. Repeat these steps until the largest F value is no longer $> F_{\mathrm{IN}}$, or all variables are added.

Backward Elimination

Begin with all K candidate regressors. Then compute the partial F-statistic for each regressor as if it were the last one to enter the model. The smallest of these partial F-statistics is compared with a preselected F-value, F_{OUT}. If the smallest partial F-value is less than F_{OUT}, remove that regressor, and refit the model. Calculate new partial F-statistic, and repeat this process. Stop when the smallest partial F value is not less than the preselected cutoff value, F_{OUT}.

Stepwise Regression

In each step, all regressors entered into the model thus far are reassessed with their partial F statistics to see if it has become redundant. If the F statistic is less than F_{OUT}, then it is removed. Generally $F_{\mathrm{IN}} > F_{\mathrm{OUT}}$ so it makes it harder to add variables than to remove them.

Logistic Regression

Logistic distribution

$$f(x) = \frac{e^x}{(1+e^x)^2}$$

Cumulative distribution function

$$F(t) = \frac{e^t}{1+e^t}$$

$E(X) = 0$, $\mathrm{Var}(X) = \pi/3$. Suppose Y is a binary response variable,

$$Y_i = \begin{cases} 1 & \beta_0^* + \beta_1^* X_i + \epsilon_i^* < k \\ 0 & \beta_0^* + \beta_1^* X_i + \epsilon_i^* > k \end{cases}$$

$\pi_i = P(Y_i = 1) = \frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}}$. $\beta_0 = k - \beta_0^*$, $\beta_1 = -\beta_1^*$.

Estimating Parameters

Log-odds is defined as

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$$

Estimates for β_0, β_1 must be obtained numerically,

$$\hat{\pi} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)}$$

The ods ratio at a point X_0 is defined as

$$\hat{O}_R = \frac{\mathrm{odds}_{X_0+1}}{\mathrm{odds}_{X_0}} = e^{\hat{\beta}_1}$$

With repeat observations, $Y_i \sim \mathrm{Bin}(n_i, \pi_i)$,

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n}\binom{n_i}{Y_i}\pi_i^{Y_i}(1-\pi_i)^{n_i - Y_i}$$

For multiple linear regression, $X_i'\beta = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{1,p-1}X_{i,p-1}$, $E(Y) = \frac{\beta' X}{1+\beta' X}$, $log\frac{\pi}{1-\pi} = \beta' X$.

Testing on Coefficients

To test if several coefficients are 0, use

$$G^2 = -2\ln\left(\frac{L(RM)}{L(FM)}\right) = 2\ln\left(\frac{L(FM)}{L(RM)}\right)$$

$$\ln L(FM) = \sum_{i=1}^{n} y_i \ln\hat{\pi}_i + \sum_{i=1}^{n}(n_i - y_i)\ln(1-\hat{\pi}_i)$$

$$\ln L(RM) = y\ln y + (n-y)\ln(n-y) - n\ln n$$

We reject the null hypothesis if $G^2 > \chi^2_{p-1}$.

Goodness of Fit

We want to test $H_0 : E(Y) = \left(1+e^{-X'\beta}\right)^{-1}$. Use Pearson chi-square statistic,reject when $\chi^2 > \chi^2_{n-p}$.

$$\chi^2 = \sum_{i=1}^{n}\frac{y_i - n_i\hat{\pi}_i}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

More on Testing

Homser-Lemenshow Test

The Hosmer-Lemenshow statistic is Pearson chi-square goodness-of-fit statistic comparing observed and expected frequences, and is given as

$$HL = \sum_{j=1}^{j}\frac{(O_j - N_j\hat{\pi}_j)^2}{N_j\hat{\pi}_j(1-\hat{\pi}_j)}$$

If the fitted model is correct, $HL \sim \chi^2_{g-1}$. Reject for large values of HL.

Deviance Goodness of Fit

Uses likelihood ratio to compare reduced model $E(Y_i) = \left(1+e^{-X_i'\beta}\right)^{-1}$ and full model $E(Y_i) = \pi_i$,

$\mathrm{Dev}(X_0, X_1, \ldots, X_{p-1}) = -2(\ln L(RM) - \ln(FM))$

We reject when $\mathrm{Dev} > \chi^2_{n-p}$.

Diagonistics Measures for Logistic Regression

The residuals are defined as $e_i = Y_i - \hat{\pi}_i$, these do not have constant variance. The deviance residuals are

$$d_i = \pm\left\{2\left[Y_i\ln\left(\frac{Y_i}{n_i\hat{\pi}_i}\right) + (n_i - Y_i)\ln\left(\frac{n_i - Y_i}{n_i(1-\hat{\pi}_i)}\right)\right]\right\}^{1/2}$$

The standardized Pearson residuals as $r_i = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$ which do not have unit variance. The studentized deviance Pearson residuals are

$$sr_i = \frac{r_i}{\sqrt{1-h_{ii}}}$$

h_{ii} is the ith diagonal entry of the hat matrix, $H = V^{1/2}X(X'VX)^{-1}V^{1/2}$. V is the diagonal matrix with $V_{ii} = n_i\hat{\pi}_i(1-\hat{\pi}_i)$. For an adequate model, $E(Y_i) = \hat{\pi}_i$, and the plots of sr_i vs $\hat{\pi}_i$ and sr_i vs linear predictors $X_i'\beta$ should show a smooth horizontal Lowess line through 0. Same for a plot of d_i vs $\hat{\pi}_i$ and d_i vs $X_i'\beta$.

Detecting Influential Cases

Delete one observation at a time and measuring its effects on the χ^2 and the Dev statistic. Plot these vs i, and look for spikes which indicate influential observations. Similarly, we can plot these vs $\hat{\pi}_i$.

Poisson Regression and GLM's

Poission regression uses Poisson distribution, $f(y) = \frac{e^{-\mu}\mu^y}{y!}$. The model is $Y_i = \mu_i + \epsilon_i$, $\mu_i = e^{X_i'\beta}$. For GLM's, response is assumed to have some exponential distributio, $\mu = E(Y) = \frac{db(\theta_i)}{d\theta_i}$, and $\mathrm{Var}(Y) = \frac{d^2b(\theta_i)}{d\theta_i^2}a(\phi)$

$$f(y_i, \theta_i, \phi) = \exp\left(\frac{y_i\theta - b(\theta_i)}{a(\phi)} + h(y_i, \phi)\right)$$