5.1 → Introduction

→ Hypothesis tests are a statistical modelling framework performed by comparing how well two related models fit the data

→ For GLMs, models must have the same probability distribution & the same link function

→ But the linear components can differ, specifically the simple model is a subset of the more general model

→ Summary stats are used to compare how well models fit the data.
Goodness of fit statistics may be based on max value of the (log)-likelihood function, minimum value of SSE or a composite statistic based on residuals

→ Summary of process & logic of hypothesis tests

1) Specify model M_0 (corresponding to H_0)
— — — — M_1 — — — — — — H_1

2) Fit M_0 & calculate goodness of fit statistic G_0
— M_1 — — — — — — — — G_1

3) Calculate the improvement in fit, usually $G_1 - G_0$ or G_1/G_0

4) Use sampling dist of $G_1 - G_0$ (or whatever from 3) to test hypothesis $G_0 = G_1$ vs $G_0 \neq G_1$

5) If $G_0 = G_1$ is not rejected ⇒ fail to reject H_0 & M_0 is preferred model
$G_0 \neq G_1$ is rejected ⇒ reject H_0 & M_1 — — — — — — — —

→ Sampling distributions

→ For both forms of inference (CI & HT), the sampling distributions are required
CI ↙ need samp dist of estimator
HT ↘ need samp dist of goodness of fit statistic

→ If response ~ normal dist, then samp dist can be determined exactly.
For others, need asymptotics based on CLT
(ignoring attention to regularity conditions
⇒ If Y_i are ⫫ & from exp family ⇒ conditions are met)

→ Basic idea is that under appropriate conditions if S is a statistic of interest then approximately

$$\frac{S - E(S)}{\sqrt{V(S)}} \sim N(0,1)$$

or equivalently

$$\frac{(S - E(S))^2}{V(S)} \sim \chi^2_1$$

→ If there is a vector of statistics $S = \begin{bmatrix} S_1 \\ \vdots \\ S_p \end{bmatrix}$ w/ asymptotic expectation $E(S)$ & var-cov matrix V

$$[S - E(S)]^T V^{-1} [S - E(S)] \sim \chi^2_p$$

↳ nonsingular & inverse exists

5.2 → Sampling dist for score statistics

→ setup → Suppose $Y_1, \dots, Y_N$ are ⫫ RVs in a GLM w/ parameters β, where $E(Y_i) = \mu_i$, $g(\mu_i) = x_i^T \beta = \eta_i$

→ Score statistic → $U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{N} \left[\frac{Y_i - \mu_i}{V(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right]$ for $j = 1, \dots, p$

→ As $E(Y_i) = \mu_i$ for all i ⇒ $E(U_j) = 0$ for $j = 1, \dots, p$

→ Var-cov matrix of U_j is the information matrix J w/ elements $J_{jk} = E[U_j U_k]$

→ If only one parameter β, score stat has asymptotic dist

$$\frac{U}{\sqrt{J}} \sim N(0,1) \quad \text{or equivalently} \quad \frac{U^2}{J} \sim \chi^2(1)$$

$(E(U) = 0 \ \& \ V(U) = J)$

→ If there is a vector of parameters $\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$ & score vector $U = \begin{bmatrix} U_1 \\ \vdots \\ U_p \end{bmatrix} \sim MVN(0, J)$ (asymptotically)

⇒ $U^T J^{-1} U \sim \chi^2(p)$ for large samples

→ Example → $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$
↳ known constant

→ log likelihood → $l = \frac{-1}{2\sigma^2} \sum (Y_i - \mu)^2 - N \ln(2\pi\sigma^2)$

→ Score statistic → $U = \frac{dl}{d\mu} = \underset{①}{\frac{1}{\sigma^2} \sum (Y_i - \mu)} = \underset{②}{\frac{N}{\sigma^2} (\bar{Y} - \mu)}$

→ MLE → obtained by solving $U = 0$ ⇒ $\hat{\mu} = \bar{Y}$

→ Expected value & variance → $E(U) = \langle \text{using } ① \rangle = \frac{1}{\sigma^2} \sum [\underbrace{E(Y_i)}_{=\mu} - \mu] = 0$

→ $J = V(U) = \langle \text{using } ① \rangle = \frac{1}{\sigma^4} \sum V(Y_i) = \frac{N}{\sigma^2}$

⇒ $\frac{U}{\sqrt{J}} = \langle \text{using } ② \rangle = \frac{(\bar{Y} - \mu)}{\sigma / \sqrt{N}} \sim N(0,1)$

asymptotically (exactly in this scenario b/c $\bar{Y} \sim N(\mu, \sigma^2/N)$)

Similarly $U^T J^{-1} U = \frac{U^2}{J} = \frac{(\bar{Y} - \mu)^2}{\sigma^2 / N} \sim \chi^2_1$

⇒ Use this sampling dist of U to make inferences about μ
e.g.) 95% CI = $\bar{y} \pm 1.96 \, \sigma/\sqrt{N}$

→ Example → If $Y \sim \text{Bin}(n, \pi)$

→ log-likelihood → $l(\pi | y) = y \ln(\pi) + (n - y) \ln(1 - \pi) + \ln \binom{n}{y}$

→ Score statistic → $U = \frac{dl}{d\pi} = \frac{Y}{\pi} - \frac{n - Y}{1 - \pi} = \frac{Y - n\pi}{\pi(1 - \pi)}$

→ Mean & variance → $E(Y) = n\pi$ ⇒ $E(U) = 0$ as expected

→ $V(Y) = n\pi(1 - \pi)$ ⇒ $V(U) = \frac{1}{\pi^2 (1 - \pi)^2} V(Y) = \frac{n}{\pi(1 - \pi)}$
↳ just substitute $U = f(Y)$

⇒ $\frac{U}{\sqrt{J}} = \frac{Y - n\pi}{\pi(1 - \pi)} \Big/ \sqrt{\frac{n}{\pi(1 - \pi)}} = \frac{Y - n\pi}{\sqrt{n\pi(1 - \pi)}} \sim N(0,1)$

↳ normal approximation to the binomial used for CI & tests about π

5.3 → Taylor series approximations

→ Taylor series approximations
→ used to obtain asymptotic sampling distributions for various other statistics

→ $f(x) = f(t) + (x - t) \left[\frac{df}{dx} \right]_{x=t} + \frac{1}{2} (x - t)^2 \left[\frac{d^2 f}{dx^2} \right]_{x=t} + \dots$

provided x is near t

→ For log-likelihood function of single parameter β, the first 3 terms become

$$l(\beta) = l(b) + (\beta - b) U(b) + \frac{1}{2} (\beta - b)^2 U'(b)$$

↳ score function evaluated at β

$U' = \frac{d^2 l}{d\beta^2} \approx E(U') = -J$ ⟨from earlier formulas⟩

⇒ approximation

$$l(\beta) = l(b) + (\beta - b) U(b) - \frac{1}{2} (\beta - b)^2 J(b)$$

↳ information evaluated at $\beta = b$

→ For vector β, this becomes

$$l(\beta) = l(b) + (\beta - b)^T U(b) - \frac{1}{2} (\beta - b)^T J(b) (\beta - b)$$

→ For score function of a single parameter β, using first two terms of series, we get

$U(\beta) = U(b) + (\beta - b) \underbrace{U'(b)}_{\approx E(U') = -J}$

$= U(b) - (\beta - b) J(b)$

for vector ⇒ $U(\beta) = U(b) - J(b)(\beta - b)$ ✩

5.4 → Sampling distributions for MLEs

→ ✩ equation directly above can be used to find sampling dist of MLE $b = \hat{\beta}$. By def, b is the estimator which maximizes $l(b)$ & so $U(b) = 0$ (first derivative)

⇒ $U(\beta) = -J(b)(\beta - b)$ or equivalently
$(b - \beta) = J^{-1} U$ ⟨left multiply J^{-1} & use negative to switch b & β⟩

→ If J is treated as constant
$E(b - \beta) = 0$ b/c $E(U) = 0$
⇒ $E(b) = \beta$ ⇒ b is a consistent estimator of β
(at least asymptotically)

→ The var-cov matrix for b is

$E[(b - \beta)(b - \beta)^T] = E[(J^{-1} U)(J^{-1} U)^T]$ ⟨like $\text{Cov}(X, Y)$: $E((X - \mu_X)(Y - \mu_Y)^T)$⟩

$= J^{-1} \underbrace{E(U U^T)}_{= J} J^{-1}$ constant & $(J^{-1})^T = J^{-1}$ b/c symmetric

$= J^{-1}$

⇒ Asymptotic sampling dist for b is

$(b - \beta)^T (J(b)^{-1})^{-1} (b - \beta) =$
$(b - \beta)^T J(b) (b - \beta) \sim \chi^2(p)$

→ This is the Wald statistic

→ for one parameter → more commonly used form → $b \sim N(\beta, J^{-1})$

→ If response ~ normal ⇒ is exact result

→ Example → Consider model $E(Y_i) = \mu_i = x_i^T \beta$, $Y_i \overset{⫫}{\sim} N(\mu_i, \sigma^2)$
⇒ glm w/ identity link function

→ Information matrix → $J_{jk} = \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{V(Y_i)} \underbrace{\left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2}_{= 1}$

$= \sum \frac{x_{ij} x_{ik}}{\sigma^2}$

$= \frac{1}{\sigma^2} X^T X$

→ $z_i = \underbrace{\sum_{k=1}^{p} x_{ik} b_k^{(m-1)}}_{= x_i^T \beta} + \underbrace{(y_i - \mu_i)}_{\text{evaluated @ } b^{(m-1)}} = y_i$
$\mu_i = x_i^T \beta$

⇒ $X^T W X b^{(m)} = X^T W z$ is equal to

$\frac{1}{\sigma^2} X^T X b = \frac{1}{\sigma^2} X^T y$

acting like W ⇒ MLE is $b = (X^T X)^{-1} X^T y$

→ writing model in matrix form → $Y \sim MVN(X\beta, \sigma^2 I)$

→ expected value & variance of b

→ $E(b) = E[(X^T X)^{-1} X^T Y]$
$= (X^T X)^{-1} X^T \underbrace{E(Y)}_{= X\beta}$
$= (X^T X)^{-1} X^T X \beta$
$= \beta$ ⇒ unbiased

→ Var-cov matrix of b

$b - \beta = (X^T X)^{-1} X^T Y - \beta$
$= (X^T X)^{-1} X^T Y - \underbrace{(X^T X)^{-1} (X^T X)}_{= 1} \beta$
$= (X^T X)^{-1} X^T (Y - X\beta)$

$(AB)^T = B^T A^T$

⇒ $E[(b - \beta)(b - \beta)^T] = (X^T X)^{-1} X^T E[(Y - X\beta)(Y - X\beta)^T] X (X^T X)^{-1}$
$= (X^T X)^{-1} X^T V(Y) X (X^T X)^{-1}$
$= \sigma^2 (X^T X)^{-1}$
$= J^{-1}$

→ In end → MLE b is a linear combo of Y_i & b/c $Y_i \sim$ Normal
$b \sim$ normal ⇒ exact sampling dist
$b \sim N(\beta, J^{-1})$, or equivalently
$(b - \beta)^T J (b - \beta) \sim \chi^2(p)$

5.5 → Log-likelihood ratio statistic

→ One way of assessing the adequacy of a model is to compare it with a more general model w/ the max # of parameters that can be estimated.
This is called a saturated (maximal or full) model

→ If there are N obs $Y_1, \dots, Y_N$ all w/ potentially different values of the linear component $x_i^T \beta$, then a saturated model can be specified w/ N parameters

→ If there are replicates (obs w/ same combos of X), then # parameters in full model may be $< N$

→ setup → Let m be the max # of parameters that can be estimated
& let β_{max} denote the parameter vector for the saturated model
& $b_{max} = \hat{\beta}_{max\ MLE}$

→ The likelihood & log-likelihood functions evaluated at b_{max} will be larger than any other likelihood function (using same dist & link function)
b/c it provides the most complete description of the data

→ Statistic → Let $L(b | y)$ be the likelihood function for the model of interest then the likelihood ratio

$$\lambda = \frac{L(b_{max} | y)}{L(b | y)} \Rightarrow \ln(\lambda) = l(b_{max} | y) - l(b | y)$$

provides a way of assessing the goodness of fit for the model

→ large values of λ or $\ln(\lambda)$ ⇒ poor description for model of interest relative to saturated model

→ Need sampling dist of $\ln(\lambda)$ to determine critical values

→ But $\underbrace{2 \ln(\lambda)}_{\text{Deviance}} \sim \chi^2$ ⇒ use this statistic instead

5.6 → Sampling dist of deviance

→ Definition → Deviance (log-likelihood (ratio) statistic)

$$D = 2 [l(b_{max} | y) - l(b | y)]$$

→ Distribution → If b is MLE of β (so that $U(b) = 0$)

(earlier equation) Taylor series approximation $l(\beta) = l(b) + (\beta - b)^T \underbrace{U(b)}_{= 0} - \frac{1}{2} (\beta - b)^T J(b) (\beta - b)$

$l(\beta) - l(b) = -\frac{1}{2} (\beta - b)^T J(b) (\beta - b)$

approximately

⇒ $2 [l(b) - l(\beta)] = (\beta - b)^T J(b) (\beta - b) \sim \chi^2(p)$
↳ # of parameters
⟨order switched from earlier derivation, but essentially squaring ⇒ not important⟩

→ Sampling dist of D

→ $D = 2 [l(b_{max}) - l(b)]$

$= \underbrace{2 [l(b_{max}) - l(\beta_{max})]}_{\sim \chi^2_m} - \underbrace{2 [l(b) - l(\beta)]}_{\sim \chi^2_p} + \underbrace{2 [l(\beta_{max}) - l(\beta)]}_{= \nu}$

$\sim \chi^2_m$ ↳ # params in saturated model

$\sim \chi^2_p$ ↳ # params in model of interest

$= \nu$ → positive constant near zero if model of interest fits data almost as well as the saturated model fits

⇒ $D \sim \chi^2(m - p, \nu)$
↳ ncp

→ Use in tests

→ D is the TS in most HT for GLMs

→ If Y_i's ~ Normal ⇒ $D \sim \chi^2$ exactly

→ however it depends on $V(Y_i) = \sigma^2$ which is usually unknown
⇒ D cannot be used directly as a GOF stat

→ for other dists → $D \overset{approx}{\sim} \chi^2$

→ however for Binomial & Poisson, D can be calculated & used directly as a GOF stat

→ Example → Let $Y_i \overset{⫫}{\sim} \text{Bin}(n_i, \pi_i)$

→ Log-likelihood → $l(\beta) = \sum_{i=1}^{N} \left[y_i \ln(\pi_i) - y_i \ln(1 - \pi_i) + n_i \ln(1 - \pi_i) + \ln \binom{n_i}{y_i} \right]$

→ full model → MLE → for a saturated model, all π_i's are different, so $\beta = [\pi_1, \dots, \pi_N]^T$

→ MLE for each is $\hat{\pi}_i = y_i / n_i$

→ likelihood → max value of the log-likelihood function is

$$l(b_{max} | y) = \sum \left[y_i \ln\left(\frac{y_i}{n_i} \right) - y_i \ln\left(\frac{n_i - y_i}{n_i} \right) + n_i \ln\left(\frac{n_i - y_i}{n_i} \right) + \ln \binom{n_i}{y_i} \right]$$

→ Model of interest → MLE → for any other model w/ $p < N$ parameters, let $\hat{\pi}_i$ be the MLE for probabilities & let $\hat{y}_i = n_i \hat{\pi}_i$ be the fitted values

→ likelihood → the log likelihood evaluated at these values is

$$l(b | y) = \sum \left[y_i \ln\left(\frac{\hat{y}_i}{n_i} \right) - y_i \ln\left(\frac{n_i - \hat{y}_i}{n_i} \right) + n_i \ln\left(\frac{n_i - \hat{y}_i}{n_i} \right) + \ln \binom{n_i}{y_i} \right]$$

→ Final deviance → $D = 2 [l(b_{max}) - l(b)]$

$= 2 \sum_{i=1}^{N} \left[y_i \ln\left(\frac{y_i}{\hat{y}_i} \right) + (n_i - y_i) \ln\left(\frac{n_i - y_i}{n_i - \hat{y}_i} \right) \right]$

→ Example → normal model → $E(Y_i) = \mu_i = x_i^T \beta$, $Y_i \overset{⫫}{\sim} N(\mu_i, \sigma^2)$, $i = 1, \dots, N$

→ likelihood → $l(\beta | y) = -\frac{1}{2\sigma^2} \sum (y_i - \mu_i)^2 - \frac{1}{2} N \ln(2\pi\sigma^2)$

→ full model → all μ_i's can be different ⇒ $\dim(\beta) = N \times 1$

→ maximizing above likelihood gives $\hat{\mu}_i = y_i$

⇒ max value of likelihood is

$$l(b_{max} | y) = -\frac{1}{2} N \ln(2\pi\sigma^2)$$

→ Model of interest → Let $b = (X^T X)^{-1} X^T y$ be the MLE

→ The corresponding max value of the likelihood is

$$l(b | y) = -\frac{1}{2\sigma^2} \sum (y_i - x_i^T b)^2 - \frac{1}{2} N \ln(2\pi\sigma^2)$$

→ final deviance → $D = 2 [l(b_{max}) - l(b)]$

$= \frac{1}{\sigma^2} \sum (y_i - x_i^T b)^2$

$= \frac{1}{\sigma^2} \sum (y_i - \hat{\mu}_i)^2$

→ If only one parameter → $E(Y_i) = \mu$ for all i, $X = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$, $b = \hat{\mu} = \bar{y}$ & $\hat{\mu}_i = \bar{y}$ for all i

→ $D = \frac{1}{\sigma^2} \sum (y_i - \bar{y})^2$

→ This is related to the sample variance S^2

$$S^2 = \frac{1}{N-1} \sum (y_i - \bar{y})^2 = \frac{\sigma^2 D}{N - 1}$$

→ we know $\frac{N-1}{\sigma^2} S^2 \sim \chi^2_{N-1}$ ⇒ $D \sim \chi^2_{N-1}$ exactly

→ More generally → $D = \frac{1}{\sigma^2} \sum (y_i - x_i^T b)^2$

$= \frac{1}{\sigma^2} \underbrace{(y - Xb)}^T (y - Xb)$

$= y - X (X^T X)^{-1} X^T y$
$= [I - X (X^T X)^{-1} X^T] y$
$= (I - H) y$
hat matrix $= X (X^T X)^{-1} X^T$

⇒ quadratic form → $(y - Xb)^T (y - Xb) = \{(I - H) y\}^T \{(I - H) y\}$

$= y^T (I - H) y$ ⟨b/c H is idempotent → $HH = H$ & $H^T = H$⟩
rank $= n$, rank $= p$
rank$(I - H) = n - p$

⇒ $D \sim \chi^2(n - p, \lambda)$
↳ $\lambda = (X\beta)^T (I - H) (X\beta) / \sigma^2$

→ but $(I - H) X = 0$ ⇒ $D \sim \chi^2_{N-p}$ exactly

→ Scaled deviance → $\sigma^2 D = \sum (y_i - \hat{\mu}_i)^2$

→ If model fits the data well, then $D \sim \chi^2_{N-p}$.
The expected value for a RV with χ^2 dist is $N - p$

⇒ $E(D) = N - p$

→ This provides an estimate for σ^2 b/c

$$\hat{\sigma}^2 = \frac{\sum (y_i - \hat{\mu}_i)^2}{N - p}$$

↳ for normal linear models, some softwares output scaled deviance & call $\hat{\sigma}^2$ called the scale parameter

→ Deviance is also related to sum of squares of standardized residuals

$$\sum r_i^2 = \frac{1}{\hat{\sigma}^2} \sum (y_i - \hat{\mu}_i)^2$$

→ provides rough rule of thumb for the overall magnitude of the standardized residuals

→ If the model fits the data well so that $D \sim \chi^2_{N-p}$, you could expect $\sum r_i^2 \approx N - p$

→ Example → Poisson model → $Y_i \overset{⫫}{\sim} \text{Poisson}(\lambda_i)$

→ likelihood → $l(\beta | y) = \sum y_i \ln(\lambda_i) - \sum \lambda_i - \sum \ln(y_i!)$

→ full model → all λ_i's are different ⇒ $\beta = [\lambda_1, \dots, \lambda_N]^T$

→ MLE → $\hat{\lambda}_i = y_i$

→ max value → $l(b_{max}) = \sum y_i \ln(y_i) - \sum y_i - \sum \ln(y_i!)$

→ Model of interest → $p < N$ parameters

→ MLE → b can be used to calculate estimates $\hat{\lambda}_i$

⇒ fitted values $\hat{y}_i = \hat{\lambda}_i$ b/c $E(Y_i) = \lambda_i$

→ max value → $l(b) = \sum y_i \ln(\hat{y}_i) - \sum \hat{y}_i - \sum \ln(y_i!)$

→ final deviance → $D = 2 [l(b_{max}) - l(b)]$

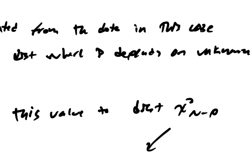
$= 2 \left[\sum y_i \ln(y_i / \hat{y}_i) - \sum (y_i - \hat{y}_i) \right]$

→ For most models, it can be shown $\sum y_i = \sum \hat{y}_i$

⇒ $D = 2 \sum o_i \ln(o_i / e_i)$
observed value y_i; estimated expected value $\hat{y}_i$

→ This value can be calculated from the data in this case (unlike for the normal dist where D depends on unknown constant σ^2)

→ Can compare this value to dist χ^2_{N-p}



⇒ Small values (relative to df) mean good fit
↳ expected value

5.7 → Hypothesis testing

→ previous approaches → Hypotheses about the parameter vector β of length p can be tested using the sampling dist of the Wald statistic $(\hat{\beta} - \beta)^T J (\hat{\beta} - \beta) \sim \chi^2_p$
Occasionally the score statistic is used $U^T J^{-1} U \sim \chi^2_p$

→ An alternative approach is to compare the GOF of two nested or hierarchical models
(same prob dist & link function, linear component M_0 is a subset of linear component M_1)

→ new approach → H_0: $\beta = \beta_0 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix}$ → M_0

H_1: $\beta = \beta_1 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$ → M_1 $\quad q < p < N$

→ We can test H_0 vs H_1 using the difference of deviance stats

$\Delta D = D_0 - D_1 = 2 [l(b_{max}) - l(b_0)] - 2 [l(b_{max}) - l(b_1)]$

$= 2 [l(b_1) - l(b_0)]$

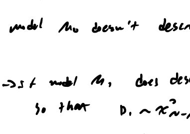
→ If both models describe the data well, then $D_0 \sim \chi^2(N - q)$ & $D_1 \sim \chi^2(N - p)$ ⇒ $\Delta D \sim \chi^2_{(p-q)}$ (provided some ⫫ conditions hold)

→ If ΔD is consistent w/ this dist, choose M_0 b/c simpler
(although keep in mind practical importance of predictors & not just the statistical significance)

→ If model M_0 doesn't describe the data well, then D_0 tend to be $>$ $E(D_0)$
↳ χ^2_{N-q}
non-central χ^2
$E(\cdot)$ >

→ If model M_1 does describe the data well so that $D_1 \sim \chi^2_{N-p}$ ⇒ if M_0 doesn't describe data well, then ΔD will be bigger than expected from $\chi^2(p-q)$

→ ⟨Usual testing conclusion⟩



→ Keep in mind model M_1 may not fit the data particularly well

→ Provided ΔD can be calculated, it is a good method for HT & sampling dist of ΔD (multiple deviances) is usually better approximated by χ^2 dist rather than using only a single deviance

→ For normal & some other dists, Deviance isn't totally determined by the data b/c of nuisance parameters that aren't estimated
But this can be overcome ↓

→ Example → HT normal model

$E(Y_i) = \mu_i = x_i^T \beta$ $\quad Y_i \overset{⫫}{\sim} N(\mu_i, \sigma^2)$

$D = \frac{1}{\sigma^2} \sum (y_i - \hat{\mu}_i)^2$

→ Have $\hat{\mu}_i(0)$ & $\hat{\mu}_i(1)$ be from M_0 & M_1, respectively

Then $D_0 = \frac{1}{\sigma^2} \sum (y_i - \hat{\mu}_i(0))^2$ &

$D_1 = \frac{1}{\sigma^2} \sum (y_i - \hat{\mu}_i(1))^2$

→ It's usual to assume that M_1 fits the data well (i.e. H_1 is correct)

⇒ $D_1 \sim \chi^2_{N-p}$. If M_0 also fits the data well, then $D_0 \sim \chi^2(N-q)$

⇒ $\Delta D \sim D_0 - D_1$

→ If M_0 does not fit the data well, (i.e. H_0 is not correct) then $\Delta D \sim$ non-central χ^2

To eliminate σ^2, the following ratio is used

$F = \frac{D_0 - D_1}{p - q} \Big/ \frac{D_1}{N - p}$

$= \frac{\left[\sum (y_i - \hat{\mu}_i(0))^2 - \sum (y_i - \hat{\mu}_i(1))^2 \right] / (p - q)}{\sum (y_i - \hat{\mu}_i(1))^2 / (N - p)}$

→ Now F can be directly calculated from the fitted values

→ If H_0 is correct ⇒ $F \sim$ Central $F(p - q, N - p)$

→ If H_0 is not correct ⇒ $F \sim$ non-central $F(\dots)$