3.1 本证明参考课程ppt的大致证明，Fisher准则如下：

$$\max_{w} J_F(w) = \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{S}_1 + \tilde{S}_2}$$

又

$$\tilde{m}_1 - \tilde{m}_2 = \frac{1}{N_1}\sum_{y_j \in Y_1} y_j - \frac{1}{N_2}\sum_{y_j \in Y_2} y_j = \frac{1}{N_1}\sum_{x \in D_1} w^T x - \frac{1}{N_2}\sum_{x \in D_2} w^T x = w^T m_1 - w^T m_2$$

$$(\tilde{m}_1 - \tilde{m}_2)^2 = |w^T(m_1 - m_2)|^2 = w^T(m_1 - m_2)(m_1 - m_2)^T w = w^T S_b w$$

$$\begin{aligned}\tilde{S}_1 + \tilde{S}_2 &= \sum_{y_j \in y_1} (y_j - \tilde{m}_1)(y_j - \tilde{m}_1)^T + \sum_{y_j \in y_2} (y_j - \tilde{m}_2)(y_j - \tilde{m}_2)^T \\ &= \sum_{x \in D_1} \left(w^T x - w^T m_1\right)\left(w^T x - w^T m_1\right)^T + \sum_{x \in D_2} \left(w^T x - w^T m_2\right)\left(w^T x - w^T m_2\right)^T \\ &= \sum_{x \in D_1} w^T (x - m_1)(x - m_1)^T w + \sum_{x \in D_2} w^T (x - m_2)(x - m_2)^T w = w^T S_w w\end{aligned}$$

于是Fisher准则可以化为：

$$J_F(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}$$

不妨令分母 $\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} = c \neq 0$, 最大化分子 $\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}$，即：

$$\begin{gathered}\max \boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w} \\ s.t. \boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} = c\end{gathered}$$

使用拉格朗日乘子法可得：$L(\boldsymbol{w}, \lambda) = \boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w} - \lambda\left(\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} - c\right)$，优化此函数得到：

Fisher线性判别最优投影方向为$w^* \propto S_w^{-1}(m_1 - m_2)$

3.2 首先解释名词含义：参考百度百科

- **误差平方和**：误差平方和是根据n个观察值拟合适当的模型后，余下未能拟合部份称为**残差**，其中y平均表示n个观察值的平均值，所有n个残差平方之和称**误差平方和。作用**：衡定模型拟合效果的好坏，SSE越小说明模型拟合的效果越好。
- **决定系数（R^2）**：对于单变量回归，即为$R^2 = \frac{\operatorname{Cov}(X,Y)}{S_X S_Y}$。对于多变量回归，$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

  其中，

  $$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
  $$SSR = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2$$

  作用：决定了模型的拟合效果，当决定系数取得最大值1的时候，说明模型拟合最好。
- **MAE 和 MAPE**：MAE是平均绝对误差，表示为$MAE = \frac{1}{n}\sum_{i-1}^{n}\left|\hat{y}_i - y_i\right|$，MAPE是平均绝对百分比误差，$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$，**作用**：用于衡量模型拟合的好坏，MAE越小拟合效果越好，当MAPE为0时，大于100%时说明模型有问题。

- **MSE 和 RMSE**：MSE是均方误差，是参数估计值与参数真值之差平方的期望值：$MSE=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2$。RMSE是均方根误差，即MSE的算术平方根：$RMSE=\sqrt{MSE}$

  作用：用于衡量模型拟合效果的好坏，相比于其他系数消除了数据个数对于大小的影响，MSE越小说明拟合效果越好。

回归系数最能体现模型的准确性。

证明：

查阅百度百科等网站可得皮尔森相关系数为：

$$r=\frac{N\sum x_iy_i-\sum x_i\sum y_i}{\sqrt{N\sum x_i^2-\left(\sum x_j\right)^2}\sqrt{N\sum y_i^2-\left(\sum y_i\right)^2}}$$

又，

$$R^2=\frac{SSR}{SST}=\frac{\sum_{i=1}^{n}\left(\hat{y}_\imath-\bar{y}\right)^2}{\sum_{\{i=1}^{n}\left(y_i-\bar{y}\right)^2}$$

代入：

$$\widehat{y_2}=\widehat{\beta_0}+\widehat{\beta_1}x_i$$
$$\bar{y}=\widehat{\beta_0}+\widehat{\beta_1\bar{x}}$$
$$\widehat{\beta_1}=\frac{N\sum x_iy_i-\sum x_i\sum y_i}{N\sum x_i^2-\left(\sum x_i\right)^2}$$

代入可得：

$$\begin{aligned}R^2&=\frac{\sum_{i=1}^{n}\left(\widehat{\beta_1}\right)^2\left(x_i-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}=\left(\frac{N\sum x_iy_i-\sum x_i\sum y_i}{N\sum x_i^2-\left(\sum x_i\right)^2}\right)^2\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}\\&=\frac{N\sum x_iy_i-\sum x_i\sum y_i}{\left(N\sum x_i^2-\left(\sum x_i\right)^2\right)\left(N\sum y_i^2-\left(\sum y_i\right)^2\right)}=r^2\end{aligned}$$

3.3 **以下证明参考了https://blog.csdn.net/perryre/article/details/53678128**

3-3: Fisher准则的结论 $w = \alpha N S_w^{-1}(m_1-m_2) \propto S_w^{-1}(m_1-m_2)$

使用最小二乘法求解MSE得到：$a^* = (Y^TY)^{-1}Y^Tb$

给定 $b = [N/N_1, N/N_2]$

w_1的样本集为：$X_1 = \{x_1^1, x_2^1 \cdots x_n^1\}$

w_2的样本集为：$X_2 = \{x_1^2, x_2^2 \cdots x_n^2\}$

增广矩阵为：$Y = \begin{bmatrix} 1 & x_1^{1T} \\ \vdots & \vdots \\ 1 & x_{N_1}^{1T} \\ -1 & -x_1^{2T} \\ \vdots & \vdots \\ -1 & -x_{N_2}^{2T} \end{bmatrix} = \begin{bmatrix} e_1 & X_1 \\ -e_2 & X_2 \end{bmatrix}$

其中 $e_1 = [1 \cdots 1]$，$e_2 = [1, \cdots, 1]$，增广权系数可以表示为 $a = \begin{bmatrix} w_0 \\ w \end{bmatrix}$

b可以写作 $b = \begin{bmatrix} N/N_1 \\ \vdots \\ N/N_1 \\ N/N_2 \\ \vdots \\ N/N_2 \end{bmatrix} = \begin{bmatrix} N/N_1 & e_1 \\ N/N_2 & e_2 \end{bmatrix}$

代入最优解 a^*：$\begin{bmatrix} e_1^T & -e_2^T \\ X_1^T & -X_2^T \end{bmatrix}\begin{bmatrix} e_1 & X_1 \\ -e_2 & X_2 \end{bmatrix}\begin{bmatrix} w_0 \\ w \end{bmatrix} = \begin{bmatrix} e_1^T & -e_2^T \\ X_1^T & -X_2^T \end{bmatrix}\begin{bmatrix} \frac{N}{N_1} & e_1 \\ \frac{N}{N_2} & e_2 \end{bmatrix}$

$\Rightarrow \begin{bmatrix} e_1^Te_1 + e_2^Te_2 & e_1^TX_1 + e_2^TX_2 \\ X_1^Te_1 + X_2^Te_2 & X_1^TX_1 + X_2^TX_2 \end{bmatrix}\begin{bmatrix} w_0 \\ w \end{bmatrix} = \begin{bmatrix} \frac{N}{N_1}e_1^Te_1 & -\frac{N}{N_2}e_2^Te_2 \\ \frac{N}{N_1}X_1^Te_1 & -\frac{N}{N_2}X_2^Te_2 \end{bmatrix}$

$S_i = \sum(x-m_i)(x-m_i)^T = \sum(xx^T - m_ix^T - xm_i^T + m_im_i^T)$

$= X_i^TX_i - 2N_im_im_i^T + N_im_im_i^T = X_i^TX_i - N_im_im_i^T$

又 $m_i = \frac{1}{N_i}\sum x \Rightarrow$ 代入原式得 $X_i^TX_i - N_im_im_i^T$

$\Rightarrow S_w = \sum_{i=1}^{2}(X_i^TX_i - N_im_im_i^T)$ 根据这个等式可以导出：$X_1^TX_1 + X_2^TX_2 = S_w + N_1m_1m_1^T + N_2m_2m_2^T$

$X_i^Te_i = N_im_i$ 重新代入上面的大矩阵：

$\begin{bmatrix} N_1+N_2 & (N_1m_1+N_2m_2)^T \\ N_1m_1+N_2m_2 & S_w + N_1m_1m_1^T + N_2m_2m_2^T \end{bmatrix}\begin{bmatrix} w_0 \\ w \end{bmatrix} = \begin{bmatrix} \frac{N}{N_1}N_1 - \frac{N}{N_2}N_2 \\ \frac{N}{N_1}N_1m_1 - \frac{N}{N_2}N_2m_2 \end{bmatrix}$

根据第一行可以算得 $Nw_0 + (N_1m_1 + N_2m_2)^Tw = 0$

$w_0 = \frac{(N_1m_1 + N_2m_2)^Tw}{N}$

根据第二行和w_0的值可以得到：

$\left(-\frac{(N_1m_1+N_2m_2)(N_1m_1+N_2m_2)^T}{N}\right) = \frac{1}{N}(-N_1^2m_1m_1^T - 2N_1N_2m_1m_2^T - N_2^2m_2m_2^T)$

分为两项：第一项：$-\frac{(N_1m_1+N_2m_2)(N_1m_1+N_2m_2)^T}{N} = \frac{1}{N}(-N_1^2m_1m_1^T - 2N_1N_2m_1m_2^T - N_2^2m_2m_2^T)$

第二项：$N_1m_1m_1^T + N_2m_2m_2^T = \frac{N}{N}(N_1m_1m_1^T + N_2m_2m_2^T)$

$= \frac{1}{N}(N_1^2m_1m_1^T + N_1N_2m_1m_1^T + N_1N_2m_1m_1^T + N_2^2m_2m_2^T)$

两项之和为：$(\frac{1}{N}S_ww + \frac{N_1N_2}{N^2}(m_1-m_2)(m_1-m_2)^Tw) = m_1 - m_2$

又 $(m_1-m_2)^Tw$ 为常数：$\frac{N_1N_2}{N^2}(m_1-m_2)(m_1-m_2)^Tw = (1-\alpha)(m_1-m_2)$

$\frac{1}{N}S_ww + (1-\alpha)(m_1-m_2) = (m_1-m_2)$

$S_ww = \alpha N(m_1-m_2)$

最后得到 $w = \alpha NS_w^{-1}(m_1-m_2)$

$w^* = S_w^{-1}(m_1-m_2)$

故方向一致，两种不同的方法

3.4 **以下证明参考了课件**

罗杰思特回归需要优化的是最大似然函数：$L(\boldsymbol{w}) = \prod_{j=1}^{N} P(y_j \mid \boldsymbol{x}_j) = \prod_{j=1}^{N} \theta(y_j\boldsymbol{w}^T\boldsymbol{x}_j)$

等价于优化：

$$E(w) = -\frac{1}{N}\ln(L(w)) = -\frac{1}{N}\ln\left(\prod_{j=1}^{N}\theta(y_jw^Tx_j)\right)$$
$$= \frac{1}{N}\sum_{j=1}^{N}\ln\left(1 + e^{-y_jw^Tx_j}\right)$$

求梯度

$$
\begin{aligned}
\nabla E &= \frac{1}{N}\sum_{j=1}^{N}\nabla\left(\ln\left(1+e^{-y_j w^T x_j}\right)\right)\\
&= \frac{1}{N}\sum_{j=1}^{N}\frac{e^{-y_j w^T x_j}}{1+e^{-y_j w^T x_j}} y_j x_j\\
&= -\frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{1+e^{y_j w^T x_j}}\right) y_j x_j
\end{aligned}
$$