应用回归分析

上海财经大学　统计与管理学院

第三章 多元回归

章节概括：

- 多元线性回归模型
- 最小二乘法
- 方差分析与假设检验
- 最大似然估计和多元正态分布

多元线性回归模型

- 均值函数

$$\mathrm{E}(Y|X_1 = x_1) = \beta_0 + \beta_1 x_1$$

$$\mathrm{E}(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\mathrm{E}(Y|X = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- 方差函数 $\mathrm{Var}(Y|X = x) = \sigma^2$

- 多元线性模型

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p+e$$

几何图

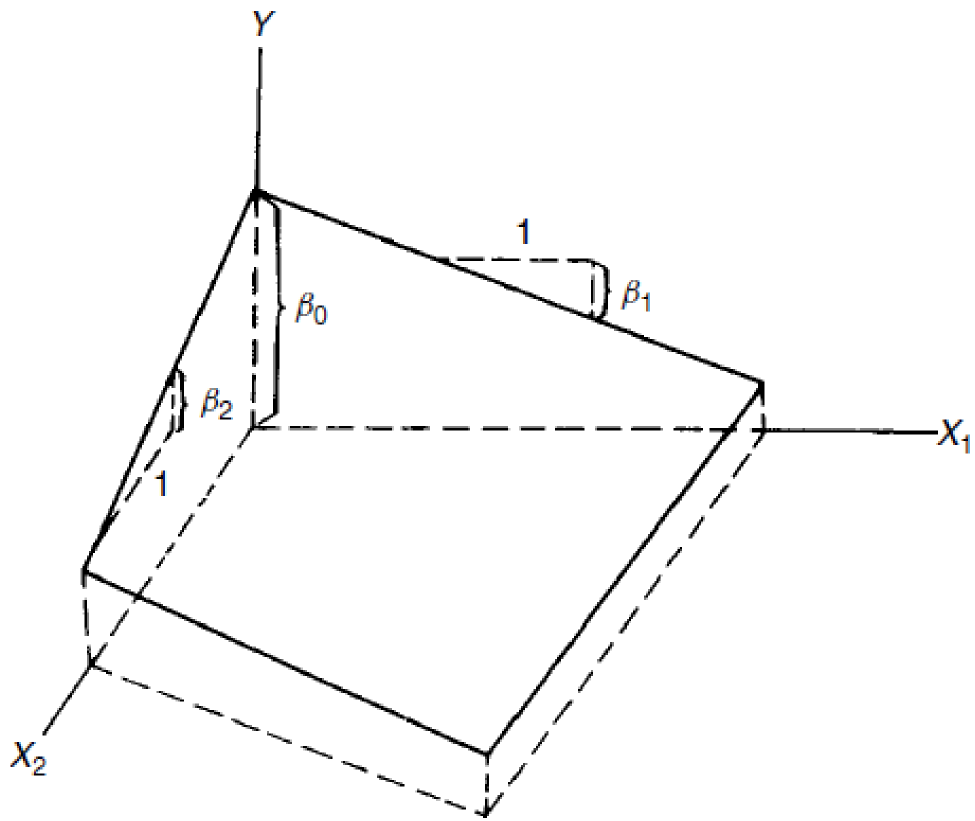

FIG. 3.2 A linear regression surface with $p = 2$ predictors.

多元线性回归模型

- n 组 独立同分布 观测数据

$$(x_{i1}, x_{i2},\ldots,x_{ip}; y_i),\ i=1,2,\ldots,n,$$

- 回归方程组

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + e_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + e_2 \\ \quad \cdots\cdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + e_n \end{cases}$$

- 线性含义：参数

矩阵表达

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \text{and} \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

$$\begin{aligned} \mathrm{E}(Y|X = \mathbf{x}_i) &= \mathbf{x}_i'\boldsymbol{\beta} \\ &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \end{aligned}$$

标准误差

- 假设1：独立、均值零，等方差

$$\mathrm{E}(\mathbf{e}) = \mathbf{0} \qquad \mathrm{Var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n$$

- 假设2： 独立正态同分布

$$\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

$$Y \sim N(\mathbf{X\beta}, \sigma^2 \mathbf{I}_n)$$

最小二乘法

- 最小二乘函数

$$RSS(\beta) = \sum (y_i - \mathbf{x}_i'\beta)^2 = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$$

- 最小二乘估计

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

- 设计矩阵X中的自变量列之间不相关, $\mathbf{X}$是一满秩矩阵

最小二乘估计

- 求解 $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2, \cdots,\ \hat{\beta}_p)$ 最小化

$$RSS(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$$

$$= \min_{\beta_0, \beta_1, \beta_2, \cdots, \beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2$$

最小二乘估计

$$
\begin{cases}
\left.\dfrac{\partial RSS}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}-\cdots-\hat{\beta}_p x_{ip})=0 \\
\left.\dfrac{\partial RSS}{\partial \beta_1}\right|_{\beta_1=\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}-\cdots-\hat{\beta}_p x_{ip})x_{i1}=0 \\
\cdots\cdots \\
\left.\dfrac{\partial Q}{\partial \beta_p}\right|_{\beta_p=\hat{\beta}_p} = -2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}-\cdots-\hat{\beta}_p x_{ip})x_{ip}=0
\end{cases}
$$

精确计算

- 中心化

$$\mathcal{X} = \begin{pmatrix} (x_{11} - \overline{x}_1) & \cdots & (x_{1p} - \overline{x}_p) \\ (x_{21} - \overline{x}_1) & \cdots & (x_{2p} - \overline{x}_p) \\ \vdots & \vdots & \vdots \\ (x_{n1} - \overline{x}_1) & \cdots & (x_{np} - \overline{x}_p) \end{pmatrix}$$

- 样本协方差阵

$$\mathcal{C} = \frac{1}{n-1} \begin{pmatrix} \mathcal{X}'\mathcal{X} & \mathcal{X}'\mathcal{Y} \\ \mathcal{Y}'\mathcal{X} & \mathcal{Y}'\mathcal{Y} \end{pmatrix}$$

- 估计

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

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}^{*\prime}\overline{\mathbf{x}}$$

回归值

- 回归值： $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}$

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

- 帽子矩阵，对称，主对角线元素记为h_{ii}

$$\mathrm{H} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' \qquad tr(H) = \sum_{i=1}^{n} h_{ii} = p+1$$

残差与方差估计

- 残差： $\hat{e} = Y - \widehat{Y} = Y - HY = (I-H)Y$

$$\operatorname{cov}(\hat{e}, \hat{e}) = \operatorname{cov}((I-H)Y, (I-H)Y)$$
$$= (I-H)\operatorname{cov}(Y, Y)(I-H) = \sigma^2 (I-H)$$

- 参差平方和： $RSS = \hat{\mathbf{e}}'\hat{\mathbf{e}} = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})$

- 方差估计： $\hat{\sigma}^2 = \dfrac{RSS}{n-(p+1)}$

估计量的统计特性

- 线性：

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

- 不相关性：

$$\operatorname{cov}(\hat{\beta}, \hat{e}) = 0$$

在正态假定下等价于独立性

估计量的统计特性

- 无偏性：

$$
\begin{aligned}
E(\hat{\beta}) &= E((X'X)^{-1}X'Y) \\
&= (X'X)^{-1}X'E(Y) \\
&= (X'X)^{-1}X'E(X\beta+e) \\
&= (X'X)^{-1}X'X\beta = \beta
\end{aligned}
$$

- 估计的方差:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

- 正态分布：

$$\hat{\beta} \sim N(\beta, (X'X)^{-1}\sigma^2)$$

估计量的统计特性

- 无偏性：

$$E(\hat{\sigma}^2) = \sigma^2$$

- 卡方分布：

$$(n-(p+1))\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-(p+1))$$

BLUE

- 高斯-马尔科夫Gauss-Markov定理

最小二乘估计量是具有最小方差的线性无偏估计量

- Best Linear Unbiased Estimator

P=1

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X}) = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \qquad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum y_i \\ \sum y_i^2 \end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{SXX} \begin{pmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

P=1

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \frac{1}{SXX} \begin{pmatrix} \sum x_i^2/n & -\overline{x} \\ -\overline{x} & \sum x_i y_i \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum y_i^2 \end{pmatrix}$$

$$= \begin{pmatrix} \overline{y} - \hat{\beta}_1 \overline{x} \\ SXY/SXX \end{pmatrix}$$

$$\mathcal{X}'\mathcal{X} = SXX \qquad \mathcal{X}'\mathcal{Y} = SXY$$

$$\hat{\beta}_1 = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathcal{Y} = \frac{SXY}{SXX}$$

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

方差分析（F检验）

- 模型的显著性检验

$$\begin{aligned}\text{NH:}\quad & \mathrm{E}(Y|X=\mathbf{x})=\beta_0\\ \text{AH:}\quad & \mathrm{E}(Y|X=\mathbf{x})=\mathbf{x}'\beta\end{aligned}$$

$$H_0:\beta_1=\beta_2=\ldots=\beta_p=0$$

- 方差分解

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

$$SSreg=SYY-RSS$$

total variation = explained + unexplained

方差分析（F检验）

TABLE 3.4 The Overall Analysis of Variance Table

Source	df	SS	MS	F	p-value
Regression	p	$SSreg$	$SSreg/1$	$MSreg/\hat{\sigma}^2$	
Residual	$n-(p+1)$	RSS	$\hat{\sigma}^2 = RSS/(n-2)$		
Total	$n-1$	SYY			

方差来源	自由度	平方和	均方	F值	P值
回归	p	$SSreg$	$SSreg/p$	$F=\frac{SSreg/p}{RSS/(n-p-1)}$	P(F>F值)=P值
残差	n-p-1	RSS	$RSS/(n\text{-}p\text{-}1)$		
总和	n-1	SYY			

$$F=\frac{SSreg/p}{RSS/(n-p-1)}$$ 当H_0成立时服从 $$F(p,n-p-1)$$

拟合优度

- 决定系数

$$R^2 = \frac{SSreg}{SYY} = 1 - \frac{RSS}{SYY}$$

单个系数检验

- 回归系数的显著性检验

$$\begin{array}{lll} \text{NH:} & \beta_1 = 0, & \beta_0, \beta_2, \beta_3, \beta_4 \text{ arbitrary} \\ \text{AH:} & \beta_1 \neq 0, & \beta_0, \beta_2, \beta_3, \beta_4 \text{ arbitrary} \end{array}$$

- 正态分布

$$\hat{\beta} \sim N(\beta, (X'X)^{-1}\sigma^2)$$

$$\hat{\beta}_i \sim N(\beta_i, c_{ii}\sigma^2)$$

T-F检验

- T检验

$$T_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{c_{jj}}\ \hat{\sigma}}$$ 当H_0成立时服从 $t(n-p-1)$

- F检验

$$\Delta SSreg_{(-j)} = SSreg(x_1,\cdots,-x_j,\cdots,x_p) - SSreg(x_1,\cdots,x_p)$$

$$F_j = \frac{\Delta SSreg_{(-j)}/1}{RSS/(n-p-1)}$$ 当H_0成立时服从 $F(1, n-p-1)$

置信区间

- T-F分布

$$t_j^2 = F_j$$

- 置信区间

β_j的置信度为1-α的置信区间为

$$(\hat{\beta}_j - t_{\alpha/2}(n-p-1)\sqrt{c_{jj}}\ \hat{\sigma},\quad \hat{\beta}_j + t_{\alpha/2}(n-p-1)\sqrt{c_{jj}}\ \hat{\sigma})$$

预测

- 原观测值

$$\hat{E}(Y|X=\mathbf{x})=\hat{y}=\mathbf{x}'\hat{\beta}$$

$$\text{sefit}(\hat{y}|\mathbf{x})=\hat{\sigma}\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}$$

- 新观测值

$$\tilde{y}_*=\mathbf{x}'_*\hat{\beta}$$

$$\text{sepred}(\tilde{y}_*|\mathbf{x}_*)=\hat{\sigma}\sqrt{1+\mathbf{x}'_*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*}$$

- 对比

$$\text{sepred}(\tilde{y}_*|\mathbf{x}_*)=\sqrt{\hat{\sigma}^2+\text{sefit}(\tilde{y}_*|\mathbf{x}_*)^2}$$

逐步方差分析

TABLE 3.5 Two Analysis of Variance Tables with Different Orders of Fitting

(a) First analysis

	Df	Sum Sq	Mean Sq
Dlic	1	86854	86854
Tax	1	19159	19159
Income	1	61408	61408
logMiles	1	34573	34573
Residuals	46	193700	4211

(b) Second analysis

	Df	Sum Sq	Mean Sq
logMiles	1	70478	70478
Income	1	49996	49996
Dlic	1	63256	63256
Tax	1	18264	18264
Residuals	46	193700	4211

最大似然估计

- Maximum Likelihood Estimation (MLE)
- 假设2：独立正态同分布

$$\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

$$\mathrm{Y} \sim \mathrm{N}(\mathbf{X\beta}, \sigma^2 \mathbf{I}_{\mathrm{n}})$$

- 一元回归

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

最大似然估计

- 似然函数

$$L(\beta_0,\beta_1,\sigma^2)=\prod_{i=1}^{n} f_i(y_i)$$

$$=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]^2\}$$

- 对数似然函数

$$\ln(L)=-\frac{n}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]^2$$

- 对比最小二乘估计

参数估计

- 估计

$$\hat{\beta}_1 = \sum \left(\frac{x_i - \bar{x}}{SXX} \right) y_i = \sum c_i y_i$$

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

- 期望与方差

$$E(\hat{\beta}_0) = \beta_0 \qquad \mathrm{Var}(\hat{\beta}_1) = \sigma^2 \frac{1}{SXX}$$

$$E(\hat{\beta}_1) = \beta_1 \qquad \mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right)$$

正态分布

- 正态分布

$$\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)\sigma^2\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{SXX}\right)$$

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{SXX}$$

方差估计

- 最小二乘法

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

- 最大似然法

$$\hat{\sigma}^2 = \frac{RSS}{n}$$

最大似然估计

- 多元正态

$$y \sim N(X\boldsymbol{\beta}, \sigma^2 I_n)$$

$$L = (2\pi)^{-n/2}\left(\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathrm{y}-\mathrm{X}\boldsymbol{\beta})'(\mathrm{y}-\mathrm{X}\boldsymbol{\beta})\right)$$

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathrm{y}-\mathrm{X}\boldsymbol{\beta})'(\mathrm{y}-\mathrm{X}\boldsymbol{\beta})$$

最大似然函数，等价于最小化 $(y-X\boldsymbol{\beta})'(y-X\boldsymbol{\beta})$，

即最小二乘法

二元正态分布

- n 组 独立同分布 观测数据

$$(x_i, y_i),\ i=1,2,\ldots,n,$$

- 二维正态分布

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right)$$

条件分布

- 给定x, y的条件分布

$$y_i | x_i \sim \mathrm{N}\left(\mu_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x_i - \mu_x), \sigma_y^2(1-\rho_{xy}^2)\right)$$

- 定义

$$\beta_0 = \mu_y - \beta_1\mu_x \qquad \beta_1 = \rho_{xy}\frac{\sigma_y}{\sigma_x}$$

$$\sigma^2 = \sigma_y^2(1-\rho_{xy}^2)$$

- 则

$$y_i | x_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

参数估计

- 二维正态参数估计

$$\hat{\mu}_x = \overline{x} \quad \hat{\sigma}_x^2 = \mathrm{SD}_x^2 \quad \hat{\rho}_{xy} = r_{xy}$$
$$\hat{\mu}_y = \overline{y} \quad \hat{\sigma}_y^2 = \mathrm{SD}_y^2$$

- 一元回归参数估计

$$\hat{\beta}_1 = r_{xy}\mathrm{SD}_y/\mathrm{SD}_x$$

$$\hat{\sigma}^2 = [(n-1)/(n-2)]\mathrm{SD}_y^2(1-r_{xy}^2)$$

多元正态分布

- 多元正态分布

$$\begin{pmatrix} \mathbf{x}_i \\ y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{xy} & \sigma_y^2 \end{pmatrix} \right)$$

- 条件分布

$$y_i | \mathbf{x}_i \sim N\left((\mu_y - \beta^{*\prime}\mu_x) + \beta^{*\prime}\mathbf{x}_i , \sigma^2 \right)$$

参数估计

- 参数估计

$$\beta^* = \Sigma_{xx}^{-1}\Sigma_{xy}$$

$$\sigma^2 = \sigma_y^2 \Sigma'_{xy}\Sigma_{xx}^{-1}\Sigma_{xy} = \sigma_y^2(1 - \mathcal{R}^2)$$

Thank You !