估计变量之间关系的强度和方向：相关性和简单线性回归

不同的术语可能用于本质上相同的模型

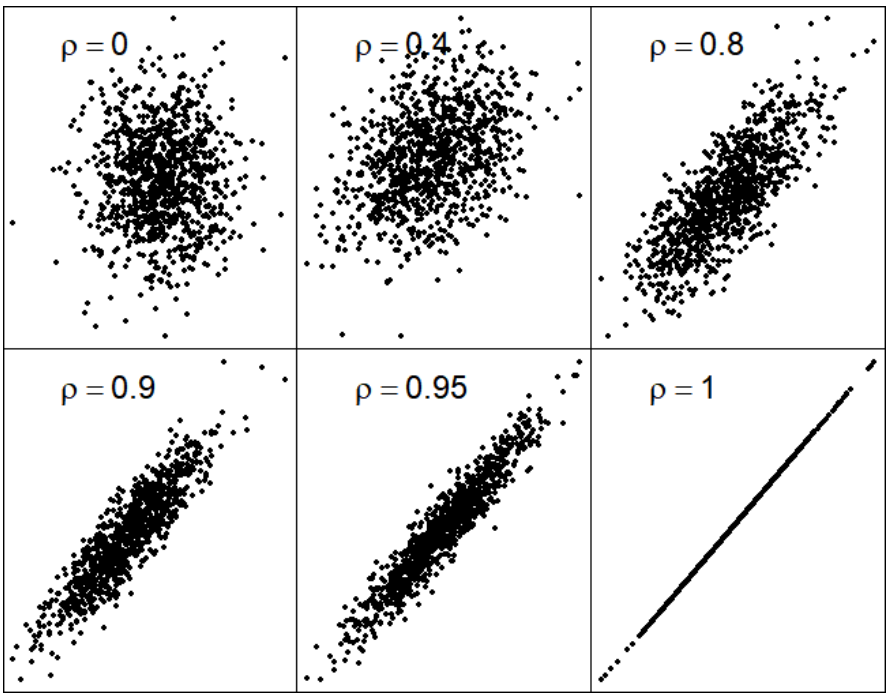
统计学家认为回归是一种通用的方法

不同的统计方法：

1. **双样本t检验（Two-sample t-test）**：
    - 1个分类解释变量，2个水平（例如，比较两组均值的差异）。
2. **方差分析（ANOVA）**：
    - 1个或多个分类解释变量，2个或多个水平（例如，比较多个组均值的差异）。
3. **协方差分析（ANCOVA）**：
    - 与ANOVA相同，但额外增加1个连续性解释变量（例如，在考虑一个连续变量的情况下比较多个组均值的差异）。
4. **线性回归（Linear Regression）**：
    - 多个连续性或分类解释变量（例如，研究多个变量如何共同影响一个连续性结果）。

相关性 (Correlation)

度量数值变量之间关系的方向和强度



从样本中估计总体参数

这张PPT讲述的是如何从样本中估计总体参数。它定义了总体参数和样本估计量的公式。具体内容如下：

总体参数：

1. 均值（Mean）：
   - μ_x 和 μ_y 分别表示变量 X 和 Y 的均值。
2. 方差（Variance）：
   - σ_x^2 和 σ_y^2 分别表示变量 X 和 Y 的方差。
3. 协方差（Covariance）：
   - σ_{xy} 表示变量 X 和 Y 的协方差。
4. 相关系数（Correlation）：
   - $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$，这是一个介于 -1 和 1 之间的数值。

样本估计量的公式：

1. 样本均值（Sample Mean）：

$$\hat{\mu}_x = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

2. 样本方差（Sample Variance）：

$$\hat{\sigma}_x^2 = s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

3. 样本协方差（Sample Covariance）：

$$\hat{\sigma}_{xy} = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

4. 样本相关系数（Sample Correlation）：

$$\hat{\rho}_{xy} = r_{xy} = \frac{s_{xy}}{s_x s_y}$$

相关性分析：

- 相关性通常指的是皮尔逊积矩相关系数，它测量 X 和 Y 之间的线性关系。

皮尔逊相关系数：

- 皮尔逊相关系数 ρ 可以通过样本相关系数 r 来估计：

$$r = \frac{s_{xy}}{s_x s_y}$$

其中，s_{xy} 是样本协方差，s_x 和 s_y 分别是 X 和 Y 的样本标准差。

假设检验：

- 相关性分析中常用的假设检验是：
  - 原假设 H_0：$\rho = 0$
  - 备择假设 H_1：$\rho \neq 0$
- 使用的检验统计量是：

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

其中，n 是样本量，t_{n-2} 表示具有 $n-2$ 个自由度的 t 分布。

Fisher变换：

- Fisher变换用于检验 ρ 是否等于任意值（包括0），以及构造 ρ 的置信区间：

$$F(r) = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

- Fisher变换的性质：
  - 虽然 $-1 \leq r \leq 1$，但 $-\infty < F(r) < \infty$

斯皮尔曼秩相关系数（Spearman's Rank Correlation Coefficient）

斯皮尔曼秩相关系数是一种通过将每个变量的数据转换为秩后计算相关性的方法。这种相关系数用于测量两个变量之间的单调关系（即随着一个变量的增加，另一个变量也增加或减少，但不要求是线性关系）。

主要特点

1. **转换为秩后的相关计算**：
   - 斯皮尔曼秩相关系数是皮尔逊相关系数的秩版本。它先将每个变量的数据转换为秩，然后再计算相关性。这意味着它关注的是数据的排序，而不是具体数值。
2. **与皮尔逊相关系数的关系**：
   - 斯皮尔曼秩相关系数与皮尔逊相关系数关系密切，但不要求线性关系。皮尔逊相关系数测量的是线性关系，而斯皮尔曼秩相关系数测量的是单调关系（即随一个变量的变化，另一个变量总是单调地变化，不要求是直线）。
3. **对异常值的稳健性**：
   - 斯皮尔曼秩相关系数对异常值（outliers）具有稳健性。这是因为它只关注数据的排序，而不是具体的数值。因此，极端值对结果的影响较小。
4. **线性关系时的相似性**：
   - 如果两个变量之间的关系是线性的且没有极端点，斯皮尔曼秩相关系数和皮尔逊相关系数的结果会非常相似。

应用场景

斯皮尔曼秩相关系数特别适用于以下情况：

- 数据中存在异常值。
- 两个变量之间的关系不是线性的，但仍希望测量它们的相关性。
- 数据是顺序的或名义的。

公式和计算方法

斯皮尔曼秩相关系数的计算公式如下：

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

其中，d_i 是第 i 个数据点的秩差，n 是数据点的数量。

秩差（rank difference）是指两个变量中对应观察值的秩之差。具体来说，对于一对观察值（X_i, Y_i），先将每个变量的所有观察值分别排序，然后计算每对观察值在排序中的差异。

以下是计算秩差的步骤：

1. **对每个变量进行排序**：将每个变量的值按大小顺序排列，并给出每个值的秩（rank）。秩是值在排序中的位置，从1开始。例如，最小值的秩为1，第二小值的秩为2，以此类推。
2. **计算秩差**：对于每对观察值（X_i, Y_i），计算两个变量对应值的秩之差，即

$$ d_i = \text{rank}(X_i) - \text{rank}(Y_i) $$

3. **计算秩差平方和**：将所有秩差的平方相加，即

$$\sum d_i^2$$

举例说明

假设我们有以下两个变量的数据：

观测编号	X	Y
1	10	20
2	20	10
3	30	30
4	40	50
5	50	40

第一步：对每个变量进行排序并计算秩

观测编号	X	Y	秩(X)	秩(Y)
1	10	20	1	2
2	20	10	2	1
3	30	30	3	3
4	40	50	4	5
5	50	40	5	4

第二步：计算秩差

观测编号	秩(X)	秩(Y)	秩差 (d_i)
1	1	2	-1
2	2	1	1
3	3	3	0
4	4	5	-1
5	5	4	1

第三步：计算秩差平方和

$$\sum d_i^2 = (-1)^2 + 1^2 + 0^2 + (-1)^2 + 1^2 = 1 + 1 + 0 + 1 + 1 = 4$$

斯皮尔曼秩相关系数的计算

使用秩差平方和，可以计算斯皮尔曼秩相关系数：

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

对于这个例子，n=5：

$$\rho = 1 - \frac{6 \times 4}{5(5^2 - 1)} = 1 - \frac{24}{120} = 1 - 0.2 = 0.8$$

斯皮尔曼秩相关系数 ρ 为 0.8，表示这两个变量之间有很强的正相关关系。

线性回归模型的前提假设

1. **线性关系（Linearity）**：
   - 假设自变量（X）和因变量（Y）之间存在线性关系，即模型形式为：
     $$Y = \beta_0 + \beta_1 X + \epsilon$$
2. **独立性（Independence）**：
   - 观测值之间的误差项是独立的，即每个 ϵ_i 互不相关。这意味着没有自相关性（例如时间序列数据中的自相关）。
3. **同方差性（Homoscedasticity）**：
   - 误差项的方差是恒定的，即对于所有 X 的值，ϵ 的方差相同。数学表达为：
     $$\text{Var}(\epsilon_i) = \sigma^2$$
4. **正态性（Normality）**：
   - 误差项服从正态分布，主要用于小样本情况下的统计推断。即：
     $$\epsilon \sim N(0, \sigma^2)$$
     这一假设确保了斜率和截距的估计量也服从正态分布，从而便于构建置信区间和进行假设检验。
5. **无多重共线性（No Multicollinearity）**：
   - 自变量之间没有完全的线性关系。对于多元线性回归模型，这意味着自变量之间的相关性不能太高，否则会导致参数估计不稳定。

验证这些假设的条件和方法

1. **线性关系的验证**:
   - 通过绘制散点图观察 X 和 Y 之间是否存在线性趋势。
   - 使用残差图（Residual Plot）检查拟合值和残差之间是否有系统性模式，若有可能存在非线性关系。
2. **独立性的验证**:
   - 对于时间序列数据，使用自相关图（ACF图）或Durbin-Watson检验来检查误差项的独立性。
3. **同方差性的验证**:

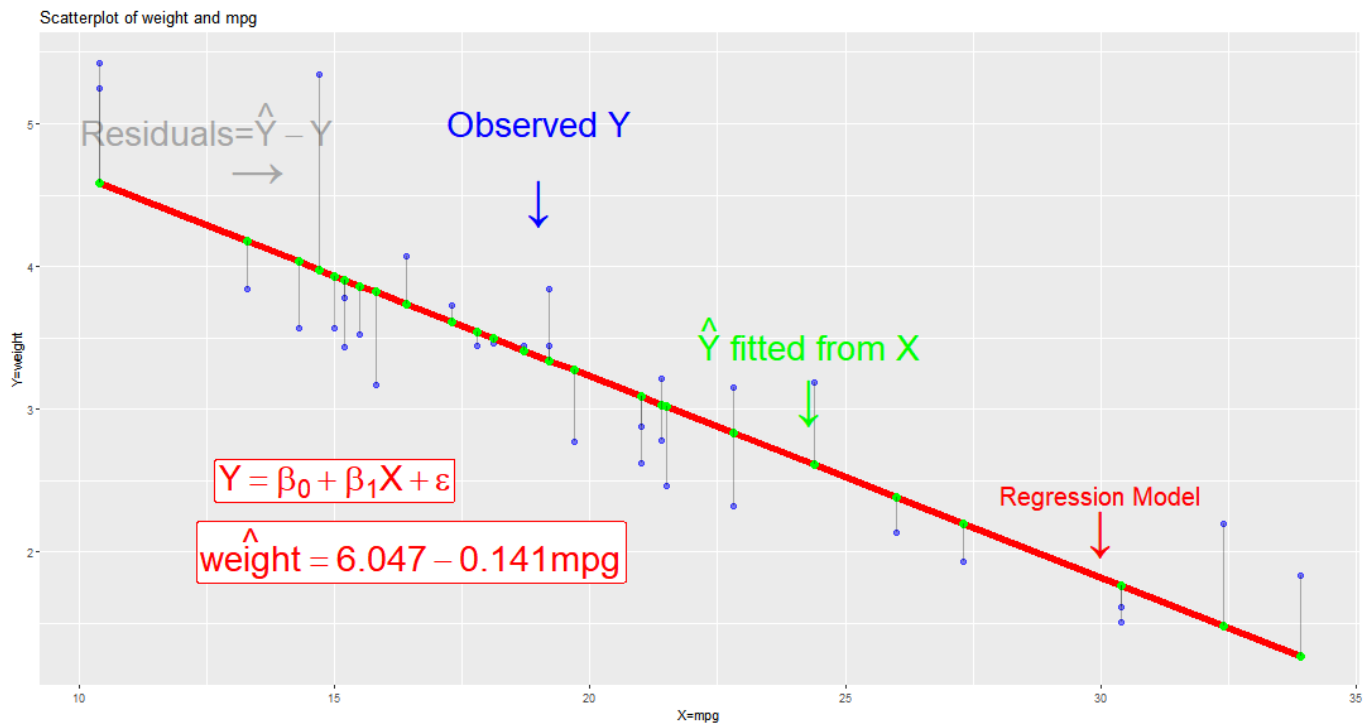
- 使用残差图检查拟合值和残差之间是否存在扇形或漏斗形状。
- 进行Breusch-Pagan检验或White检验来检测异方差性。

4. **正态性的验证**:
   - 绘制QQ图（Quantile-Quantile Plot）来检查残差的正态性。
   - 进行Shapiro-Wilk检验、Kolmogorov-Smirnov检验等正态性检验。
5. **无多重共线性的验证**:
   - 计算方差膨胀因子（Variance Inflation Factor, VIF），一般来说，VIF大于10表示存在严重的多重共线性。
   - 检查自变量之间的相关矩阵，观察是否存在高度相关的自变量对。

最佳拟合线（最小二乘法 : Least Squares)

最小化平方误差之和，该方法称为最小二乘法。 线有两个参数：

- 斜率 (Slope)
- 截距 (intercept)



模型如下

$$\hat{Y} = \beta_0 + \beta_1 X$$

$$Y = \hat{Y} + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$$

拟合值（Fitted Values）

使用估计的斜率（b_1）和截距（b_0）以及每个自变量 x_i 来计算因变量 y_i 的拟合值。

$$\hat{y}_i = b_0 + b_1 x_i$$

残差（Residuals）

残差是观察值与拟合值之间的差异。

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

残差方差的估计（Estimated Variance of the Residuals）

残差方差的估计值用于评估回归模型的拟合好坏。计算公式如下：

$$\hat{\sigma}_\epsilon^2 = s_\epsilon^2 = \frac{\sum \hat{\epsilon}_i^2}{n-2}$$

注意事项：

- 残差的平均值为0。
- 我们使用 $n-2$ 是因为我们估计了2个参数（b_0 和 b_1），所以自由度为 $n-2$

残差的最小二乘估计

最小化残差平方和（Sum of Squared Residuals, SSR）

$$SSR = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

这个过程导致残差 $\hat{\epsilon}_i$ 的和为0，因为：

1. **对截距 $\hat{\beta}_0$ 的偏导数为0**： 最小二乘估计过程要求对 $\hat{\beta}_0$ 求导，并令其等于0，这样我们得到了以下方程：

$$\sum_{i=1}^{n} \hat{\epsilon}_i = 0$$

2. **对斜率 $\hat{\beta}_1$ 的偏导数为0**： 同样地，对 $\hat{\beta}_1$ 求导并令其等于0，得到以下方程：

$$\sum_{i=1}^{n} \hat{\epsilon}_i x_i = 0$$

由于 $\sum_{i=1}^{n} \hat{\epsilon}_i = 0$，我们可以得出残差的平均值也是0：

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_i = 0$$

斜率和截距的最小二乘估计

1. **斜率的估计**：
   - 斜率（b_1）主要由 Y 和 X 之间的相关性决定。
   - 斜率的大小受 Y 和 X 标准差比率的限制。

   斜率的公式为：

   $$b_1 = r_{xy} \frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2}$$

   其中：

   - r_{xy} 是 Y和 X的相关系数。
   - s_y和 s_x分别是 Y和 X 的标准差。
   - s_{xy}是 X和 Y的协方差。
   - s_x^2 是 X 的方差。

2. **截距的估计**：
   - 通过选择回归线上一点来估计截距，然后使用斜率估计值计算截距。
   - 这点通常是 $\{\bar{x}, \bar{y}\}$，即 X 和 Y 的均值点。

   截距的公式为：

   $$b_0 = \bar{y} - b_1 \bar{x}$$

3. **斜率和截距的性质**：
   - 可以证明斜率和截距都可以看作是 Y 的加权平均值，这意味着在样本量 n 足够大的情况下，根据中心极限定理（CLT），它们都会服从正态分布。

误差项的正态分布假设

在经典的线性回归模型中，误差项的正态分布假设是非常重要的。具体来说，这个假设指的是：

$$\epsilon_i \sim N(0, \sigma^2)$$

这意味着误差项 ϵ_i 服从均值为0，方差为 σ^2 的正态分布。

为什么假设误差项为正态分布？

1. **统计推断的基础**：假设误差项为正态分布，使我们能够使用许多统计推断方法，例如计算置信区间、进行假设检验等。如果误差项不服从正态分布，这些推断的结果可能不再可靠。

2. **最小二乘估计的最佳性质**：在误差项服从正态分布的假设下，最小二乘估计量（OLS）具有最佳线性无偏估计量（BLUE）的性质。这意味着，最小二乘估计是所有线性无偏估计量中方差最小的。

3. **方便的数学性质**：正态分布具有许多方便的数学性质，使得许多推导和计算更加简单和直观。

拟合之后的误差（残差）

在实际应用中，我们通过最小二乘法拟合模型后，计算得到的残差是实际值与拟合值之间的差异：

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

这些残差并不一定严格服从正态分布，但我们希望它们近似服从正态分布。通过检查残差的分布，我们可以评估模型的适用性和假设的合理性。

如果残差不是正态分布，说明什么？

1. **模型可能不合适**：如果残差显著偏离正态分布，这可能表明模型不适合数据。例如，可能存在非线性关系，或者遗漏了重要的变量。

2. **数据可能存在异常值或不均匀方差**：残差的非正态分布也可能是由于数据中的异常值或异方差（heteroscedasticity）引起的。需要进行进一步的诊断和处理。

3. **选择其他模型**：在残差明显不服从正态分布的情况下，可以考虑使用其他更适合的模型或方法，如广义线性模型（GLM），或对数据进行转换（如对数转换）。

实际操作中的步骤

1. **拟合模型**：使用最小二乘法拟合线性回归模型，得到估计的斜率 $\hat{\beta}_1$ 和截距 $\hat{\beta}_0$

2. **计算残差**：计算每个观测值的残差 $\hat{\epsilon}_i = y_i - \hat{y}_i$

3. **检查残差的分布**：通过绘制残差直方图、QQ图等，检查残差是否近似服从正态分布。
4. **诊断和处理**：如果残差不服从正态分布，考虑是否需要对模型或数据进行调整。

斜率和截距会服从正态分布的解释

当我们说斜率和截距会服从正态分布时，我们指的是，在一定的假设下（如误差项服从正态分布，样本量足够大等），通过最小二乘法估计得到的斜率（$\hat{\beta}_1$）和截距（$\hat{\beta}_0$）的分布接近于正态分布。这是基于统计学中的中心极限定理（Central Limit Theorem, CLT）和大数定律的结果

1. **线性回归模型假设**: 在简单线性回归模型中，我们假设

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

   其中，ϵ_i 是误差项，通常假设其服从均值为 0 且方差为 σ^2 的正态分布，即

$$\epsilon_i \sim N(0, \sigma^2)$$

2. **最小二乘估计量的分布**: 最小二乘法给出了斜率和截距的估计量$\hat{\beta}_1$ 和 $\hat{\beta}_0$，它们是随机变量，因为它们依赖于随机样本。 在假设误差项 ϵ_i 服从正态分布的前提下，且样本量足够大时，$\hat{\beta}_1$ 和 $\hat{\beta}_0$ 也会近似服从正态分布。

3. **中心极限定理的作用**: 根据中心极限定理，当样本量 n 足够大时，样本均值的分布会趋向于正态分布。由于 $\hat{\beta}_1$ 和 $\hat{\beta}_0$ 是 y_i 的加权平均值，在大样本情况下，它们也会近似服从正态分布。

4. **斜率和截距的标准误**: 斜率和截距的估计量的标准误（standard errors）用于描述它们的分布。例如：

   - 斜率的标准误：

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum (x_i - \bar{x})^2}}$$

   - 截距的标准误：

$$SE(\hat{\beta}_0) = \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)}$$

正态分布性质的意义

1. **统计推断**: 由于$\hat{\beta}_1$ 和 $\hat{\beta}_0$ 近似服从正态分布，我们可以进行各种统计推断，如置信区间和假设检验。
    - **置信区间**: 对于斜率和截距的估计值，可以计算其置信区间。例如，斜率的95%置信区间为：

      $$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot SE(\hat{\beta}_1)$$

      其中，$t_{\alpha/2,n-2}$ 是t分布的临界值。
    - **假设检验**: 可以检验斜率和截距是否显著不同于某个值。例如，检验斜率是否为0（即是否存在线性关系）。
2. **模型评估**: 通过了解估计量的分布性质，我们可以更好地评估模型的稳定性和可靠性。

例子：量化汽车重量（wt）对每加仑英里数（mpg）的影响

回归模型概述

回归模型的形式是：

$$\text{mpg} = \beta_0 + \beta_1 \cdot \text{wt} + \epsilon$$

其中：

- mpg 是因变量（每加仑英里数）。
- wt 是自变量（重量）。
- β_0 是截距。
- β_1 是斜率，表示重量对每加仑英里数的影响。
- ϵ 是误差项。

表格解释

Variable	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.29	1.878	19.86	8.24e-19
wt	-5.34	0.559	-9.56	1.29e-10

R squared = 0.753

(Intercept)

- **Estimate**: 37.29
  - 截距 β_0，当重量为零时，预计的每加仑英里数是37.29。这是回归线在纵轴（mpg轴）上的截距。
- **Std. Error**: 1.878
  - 截距估计值的标准误差，表示估计值的变异程度。
- **t value**: 19.86
  - t值，用于检验截距是否显著不为零。计算公式为估计值除以标准误差。
- **Pr(>|t|)**: 8.24e-19
  - p值，表示截距显著性的概率。p值非常小（远小于0.05），表明截距显著不为零。

wt

- **Estimate**: -5.34
  - 斜率 β_1，表示重量每增加一个单位（通常是千磅），每加仑英里数预计减少5.34。这说明重量和燃油效率之间存在负相关关系。
- **Std. Error**: 0.559
  - 斜率估计值的标准误差，表示估计值的变异程度。
- **t value**: -9.56
  - t值，用于检验斜率是否显著不为零。计算公式为估计值除以标准误差。
- **Pr(>|t|)**: 1.29e-10
  - p值，表示斜率显著性的概率。p值非常小（远小于0.05），表明斜率显著不为零。

模型的 R^2 值

- R^2: 0.753
  - R^2 值表示模型解释的因变量变异的比例。值为0.753，表示模型能够解释约75.3%的每加仑英里数的变异，这表明模型拟合度较好。

R^2

R^2（决定系数）是评估回归模型拟合优度的统计量。它表示自变量（例如汽车重量）解释了因变量（例如燃油效率）总变异的比例。简单来说，R^2 告诉我们模型对数据的解释能力有多强。

R^2 的计算

R^2 的计算公式如下：

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

其中：

- RSS 是残差平方和（Residual Sum of Squares），表示模型未能解释的变异。
- TSS 是总平方和（Total Sum of Squares），表示数据中总的变异。

计算步骤如下：

1. **计算总平方和（TSS）**：

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

其中，y_i 是实际观测值，$\bar{y}$ 是因变量的平均值。

2. **计算残差平方和（RSS）**：

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

其中，$\hat{y}_i$ 是回归模型的预测值。

3. **计算 R^2**：

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

R^2 的意义

1. **解释变异**：R^2 的值在 0 到 1 之间。值越接近 1，表示模型解释了越多的因变量变异，拟合越好。值越接近 0，表示模型解释的变异很少，拟合较差。

2. **模型评价**：R^2 可用于比较不同模型的拟合优度。通常，R^2 越高，模型越优。

3. **回归模型的有效性**：高 R^2 值表明模型能够有效解释因变量的变化，但这并不一定意味着模型就是最好的。还需考虑其他指标和检验，如调整后的 R^2、AIC、BIC 以及残差分析。

调整后的 R^2 (Adjusted R^2)

调整后的 R^2 考虑了模型中自变量的数量，可以避免因增加自变量而虚增 R^2 值的情况。计算公式为：

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-k-1)}{\text{TSS}/(n-1)}$$

其中，n 是样本量，k 是自变量的数量。调整后的 R^2 更适合用于模型比较。

示例计算

假设有以下数据和回归模型：

y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
20	19	1
21	22	-1
19	20	-1
23	24	-1
22	23	-1

1. **计算总平方和（TSS）**：

$$\bar{y} = \frac{20 + 21 + 19 + 23 + 22}{5} = 21$$

$$\text{TSS} = (20-21)^2 + (21-21)^2 + (19-21)^2 + (23-21)^2 + (22-21)^2 = 1 + 0$$

2. **计算残差平方和（RSS）**：

$$\text{RSS} = 1^2 + (-1)^2 + (-1)^2 + (-1)^2 + (-1)^2 = 1 + 1 + 1 + 1 + 1 = 5$$

3. **计算 R^2**：

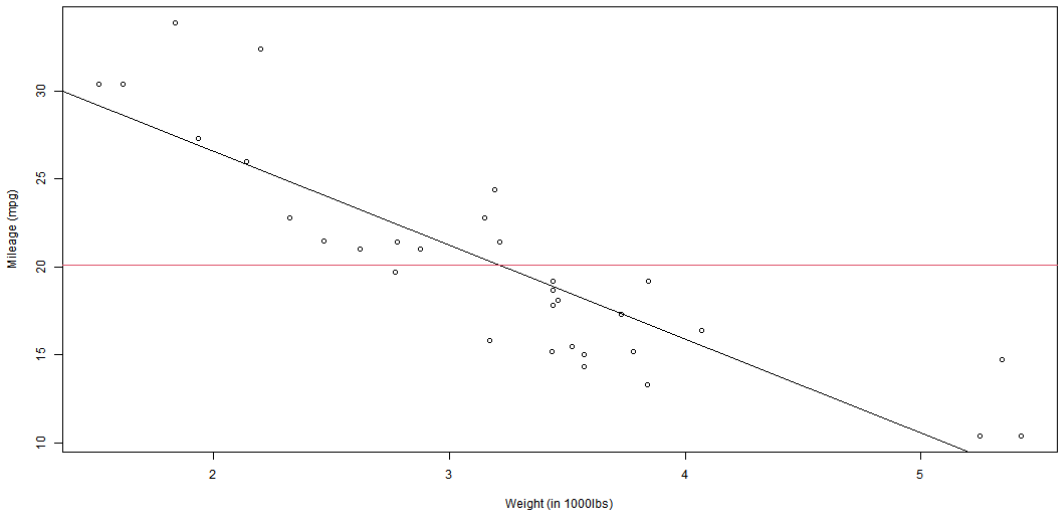
$$R^2 = 1 - \frac{5}{10} = 0.5$$

这个 R^2 值为 0.5，表示模型解释了因变量变异的 50%。

总结

- **定义**：R^2 表示回归模型解释的总变异的比例。
- **计算**：通过总平方和（TSS）和残差平方和（RSS）计算得出。
- **意义**：用于评估和比较模型的拟合优度，高 R^2 表示更好的拟合。
- **调整后的 R^2**：考虑自变量数量，提供更可靠的模型比较。

例子：与红色均值线相比，X 是否解释了 Y（带回归线）的变异性?



图像解释

图像中展示了两个回归分析中的重要元素：

1. **黑色回归线**：表示模型的拟合结果。
2. **红色均值线**：表示因变量 Y 的平均值。

图中用散点图展示了数据点及其与回归线和均值线的关系。

主要内容解释

1. **总变异（Total Variation）**：
   - 表示因变量 Y 的总变异，计算为每个实际值与均值的差值的平方和：

$$\text{TSS} = \sum (y_i - \bar{y})^2 = 1126$$

2. **残差平方和（Residual Sum of Squares, RSS）**：
   - 表示未被回归模型解释的变异，计算为每个实际值与回归预测值的差值的平方和：

$$\text{RSS} = \sum (y_i - \hat{y}_i)^2 = 278$$

3. **决定系数（R^2）**：
   - R^2 通过比较模型解释的变异与总变异来衡量模型的拟合优度：

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{278}{1126} = 0.753$$

- 这表示模型解释了因变量 75.3% 的变异。

解释的进一步详细

1. **回归模型的解释能力**：
   - R^2 值越高，说明回归模型对因变量的解释能力越强。
   - 在这个例子中，$R^2 = 0.753$，表示模型比仅使用均值来预测 Y 更好，能够解释 75.3% 的变异。
2. **决定系数的局限性**：
   - **模型复杂度**：增加模型中的变量数量通常会提高 R^2，但这也可能导致模型过拟合，从而增加预测的不确定性。
   - **样本大小的影响**：在大样本下，模型系数可能非常显著，但 R^2 仍可能较低；在小样本下，即使 R^2 较高，模型系数也可能不显著。
3. **决定系数的计算依赖于残差**：
   - R^2 依赖于模型预测值与实际值之间的误差（残差），残差越小，R^2 越大，表示模型越好。
4. **决定系数与相关系数**：
   - 对于简单线性回归，R^2 等于自变量和因变量之间的相关系数（ρ_{xy}）的平方。

拟合值误差的特性

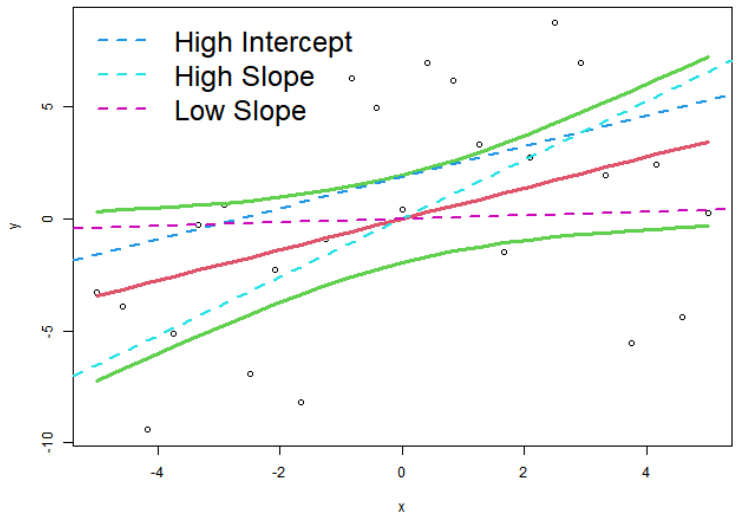
1. **误差随 x_i 接近 $\bar{x}$ 减小**：
   - 当自变量 x_i 接近其均值 $\bar{x}$ 时，拟合值的误差较小。这是因为在均值附近，模型拟合得较好。
2. **拟合值在端点处对斜率的依赖较大**：
   - 在自变量的端点处，拟合值对斜率的变化非常敏感。即，如果斜率发生微小变化，端点处的拟合值会有较大的变化。
   - 在自变量中间部分，拟合值对斜率的变化相对不敏感。
3. **误差随 x 的方差增加而减小**：
   - 自变量 x 的方差越大（即数据点在 x 轴上分布越广），拟合值的误差越小。这是因为更大的方差提供了更多的信息来准确估计斜率。

斜率的确定与端点的关系

- **斜率主要由端点决定**：
  - 由于端点处的数据点对斜率的变化敏感，斜率的估计主要受到这些数据点的影响。这意味着，端点数据点的变化会显著影响回归线的斜率。

拟合值的分布

1. **中心极限定理（CLT）的应用**：
   - 由于中心极限定理的作用，在大样本下，拟合值的分布可以近似为正态分布。这是因为拟合值是数据点的加权平均，随着样本量增加，均值分布趋近于正态分布。
2. **置信区间的计算**：
   - 由于拟合值的分布可以近似为正态分布，我们可以计算拟合值的置信区间。这意味着我们可以为拟合线上的每个点提供一个置信区间，反映估计的不确定性。



上图绿线为拟合线的 95% 的置信区间。