Linear Regression

Yingbo Li

Clemson University

MATH 9810

Linear Regression

- Response or dependent variable: Y
- Predictors or independent variables: $X_1, X_2, \ldots, X_p$

GOALS:

- Exploring $p(y|x)$ as a function of x
- Understanding the mean of Y as a function of x
- Making predictions of Y for new x.

Review: Model Assumptions

- For $i = 1, \ldots, n$,

$$Y_i = f(X_i) + \epsilon_i$$

- Regression function $E(Y \mid x) = f(x)$
- Taylors series expansion of

$$f(x_i) = f(x_0) + f'(x_0)(x_i - x_0) + \text{ Remainder}$$

leads to locally linear approximation

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- ε_i : independent errors (sampling, measurement, lack of fit)

BIG PICTURE:

Simple linear regression (one predictor plus intercept)

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

- For any x, mean of Y falls on a line: $E(Y \mid x) = \beta_0 + \beta_1 x$
- For any x, variance of Y is constant: $Var(Y \mid x) = \sigma^2$
- For any x, deviations of Y around line follow common normal distribution : $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Also can be written as $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

Estimating Regression Parameters

Preliminaries: Notations for sample summary statistics

- Sample means: $\bar{x}, \bar{y}$
- Sample variances: $s_y^2 = S_{yy}/(n-1), \quad s_x^2 = S_{xx}/(n-1)$
- Sample covariance: $s_{xy} = S_{xy}/(n-1)$
- Sums of squares are
  - $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$: Total Variation in response
  - $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$
  - $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$

Correlation

Sample correlation is covariance in a standardized scale (unit-less)

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

measure of dependence

$$-1 \leq r \leq 1$$

for a single predictor, $r^2 = R^2$ where R^2 is the coefficient of determination

Ordinary Least Squares (OLS)

For any chosen α, β,

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

measures "fit" of chosen line $\beta_0 + \beta_1 x$ to response data

- OLS estimator: Choose $\hat{\beta}_0, \hat{\beta}_1$ to *minimize* $Q(\beta_0, \beta_1)$
- Ad-hoc "principal" of least squares estimation
- Under normal error assumption OLS is equivalent to MLE

Estimating Regression Parameters

Classical approach based on maximum likelihood estimates:

$$L(\beta_0, \beta_1, \sigma^2 | Y, X) \propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i (y_i - (\beta_0 + \beta_1 x_i))^2 \right\}$$

Take derivatives and set equal to zero. We have

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Estimating Regression Parameters

Also, we get

$$\hat{\sigma}^2 = \frac{\sum_i (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{n}$$

Most software packages instead use

$$s^2_{Y|X} = \text{MSE} = \sum_{i=1}^{n} \frac{(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{n-2}$$

- Note: $n-2$ in denominator, not n or $n-1$
- Lose 2 degrees of freedom for estimation of β_0, β_1
- $s^2_{Y|X}$ is unbiased estimator of σ^2, whereas the MLE is biased.

R^2 measure of model fit:

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ (called the fitted values)
- Let SSR $= \sum_i (\hat{y}_i - \bar{y})^2$
- Let SSE $= \sum_i (y_i - \hat{y}_i)^2$

Mathematical fact that $S_{yy} =$ SSR $+$ SSE. So,

SSR$/S_{yy}$ $+$ SSE$/S_{yy}$ $= 1$

SSR$/S_{yy}$ $=$ 1 - SSE$/S_{yy}$

SSR$/S_{yy}$ is called R^2, or the coefficient of determination

Facts

R^2 is correlation squared for simple linear regression (not multiple regression)

- When model is correct, higher R^2 is better
- Measures linear correlation only
  - not general dependence
  - not causation
- Can be used to compare other simple linear regression models with transformations of X
- Does NOT provide a measure of model adequacy

Frequentist Inferences About β_1

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{V\hat{a}r(\hat{\beta}_1)}} \sim t_{n-2}(0,1)$$

- Sampling distribution of $\hat{\beta}_1$ given β_1 is t-distribution with $n-2$ degrees of freedom
- 95% confidence intervals for β_1 and tests of hypotheses (usually $H_o : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$) based on this t-distribution

Frequentist Inferences About β_0

- Sampling distribution of $\hat{\beta}_0$ is t-distribution with $n-2$ degrees of freedom

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{s^2_{Y|X}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}(0,1)$$

- Typically we care less about β_0 than about β_1

Predictions

Prediction for new case Y_{n+1} given x_{n+1}: $\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$
$\hat{Y}_{n+1}$ has t-distribution with $n-2$ degrees of freedom:

$$
\begin{aligned}
\hat{Y}_{n+1} &\sim t_{n-2}(\mu_{n+1}, s^2_{y_{n+1}}) \\
\mu_{n+1} &= \beta_0 + \beta_1 x_{n+1} \\
s^2_{y_{n+1}} &= s^2_{Y|X}\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}\right)
\end{aligned}
$$

Variance has following features:

- includes uncertainty about μ_{n+1}
- the $s^2_{Y|X}$ accounts for variation around μ_{n+1}
- increases as x_{n+1} gets further from $\bar{x}$

BIG PICTURE:

Multiple linear regression (several predictors plus intercept): here is a model for 2 predictors

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

- For any $x = (x_1, x_2)$, mean of Y falls on a line: $E(Y \mid x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- For any $x = (x_1, x_2)$, variance of Y is constant: $Var(Y \mid x) = \sigma^2$
- For any $x = (x_1, x_2)$, deviations of Y around line follow common normal distribution : $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Matrices for Multiple Regression

Write multiple regression model (with β_0 intercept) as

$$
\begin{aligned}
Y_1 &= \beta_0 + x_{11}\beta_1 + \ldots + x_{1p}\beta_p + \epsilon_1 \\
Y_2 &= \beta_0 + x_{21}\beta_1 + \ldots + x_{2p}\beta_p + \epsilon_2 \\
\vdots &= \qquad \vdots \\
Y_n &= \beta_0 + x_{n1}\beta_1 + \ldots + x_{np}\beta_p + \epsilon_n \\
\Longleftrightarrow & \\
Y &= 1\beta_0 + X_1\beta_1 + \ldots + X_p\beta_p + \epsilon \\
\Longleftrightarrow & \\
Y &= X\beta + \epsilon
\end{aligned}
$$

where $X = [1 \; X_1 \; \ldots \; X_p]$ is a $n \times (p+1)$ matrix, Y and X_j are vectors of length n, and $\beta = (\beta_0, \ldots \beta_p)$

MLEs in Matrix Notation

The MLE of β maximizes

$$Q(\beta) = (Y - X\beta)^T (Y - X\beta)$$

Equivalently, OLS solution minimizes $-Q(\beta)$.

Solution: $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $\hat{\sigma}^2 = \sum_{i=1}^n (Y_i - x_i^T \hat{\beta})^2 / n$

Most packages, including R, use $s^2_{Y|X} = \sum_{i=1}^n (Y_i - x_i^T \hat{\beta})^2 / (n - (p+1))$ rather than the MLE to estimate σ^2, because $s^2_{Y|X}$ is unbiased.

Inferences for coefficients

$$\hat{\beta} \sim t_{n-(p+1)}(\beta, (X^T X)^{-1} s^2_{Y|X}),$$

i.e., a multivariate t-distribution with $p+1$ dimensions and $n-(p+1)$ degrees of freedom.

- Components of β, say β_k, have marginal $t_{n-(p+1)}$-distributions with variance equal to the kth diagonal element of $(X^T X)^{-1} s^2_{Y|X}$
- Confidence intervals and hypothesis tests interpreted as "given all other variables are in the model, make inference for β_k"

Testing multiple coefficients

- Suppose you want to test if multiple coefficients all equal zero
- For example, you have two nested models
  - M1: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma^2)$
  - M2: $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma^2)$
- Test $H_0 : \beta_2 = \beta_3 = 0$
- Test statistic: $F = \frac{(SSE_{M2} - SSE_{M1})/(p_{M2} - p_{M1})}{SSE_{M1}/(n-(p_{M1}+1))}$
- Refer to F-distribution with $p_{M2} - p_{M1}$ degrees of freedom in numerator and $(n - (p_{M1} + 1))$ degrees of freedom in denominator
- Especially useful for sets of indicator variables.

Semi-Conjugate Priors

Regression model:

$$Y_i \overset{ind}{\sim} \mathsf{N}(\beta x_i, \sigma^2)$$

Semi-conjugate priors: independent

$$\beta \sim \mathsf{N}(b_0, \Sigma_0)$$
$$1/\sigma^2 \sim \mathsf{Gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$$

Full conditionals

$$\begin{aligned}
\beta \mid \sigma^2, Y &\sim \mathsf{N}\left(b_n, \Sigma_n\right) \\
b_n &= (\Sigma_0^{-1} + X^T X/\sigma^2)^{-1}(\Sigma_0^{-1} b_0 + X^T Y/\sigma^2) \\
\Sigma_n &= (\Sigma_0^{-1} + X^T X/\sigma^2)^{-1} \\
1/\sigma^2 \mid \beta, Y &\sim \mathsf{Gamma}((\nu_0 + n)/2, (\nu_0\sigma_0^2 + \sum_i (Y_i - \beta x_i)^2)/2)
\end{aligned}$$

Non-Informative Prior: Jeffreys Prior

Limiting case as all prior variances go to infinity and ν_0 goes to zero

$$p(\beta, \sigma^2) \propto 1/\sigma^2$$

- $\Sigma_0^{-1} = 0, \nu_0 = 0, \sigma_0^2 = 0$
- Full conditionals:

$$\beta \mid \sigma^2, Y \sim \mathsf{N}\left((X^T X)^{-1} X^T Y, (X^T X/\sigma^2)^{-1}\right)$$
$$1/\sigma^2 \mid \beta, Y \sim \mathsf{Gamma}(n/2, \sum_i (Y_i - \beta x_i)^2/2)$$

- Note that the connection with the MLE $\hat{\beta}$.

$$E(\beta \mid \sigma^2, Y) = \hat{\beta}$$
$$Var(\beta \mid \sigma^2, Y) = Var(\hat{\beta})$$

Weakly-Informative Prior: Unit Information Prior

A unit information prior is one that contains the same amount of information as that would be contained in only a single observation

- Precision of $\hat{\beta}$, i.e., its inverse variance is $X^T X/\sigma^2$, this contain the amount of information in n observations.
- Prior precision of β contain the amount of information in a single observation, $\Sigma_0^{-1} = X^T X/(n\sigma^2)$
- Prior mean $b_0 = \hat{\beta}$
  NOT a real prior distribution, because it depends on Y. But it only uses a small amount of the information in Y.
- $\nu_0 = 1, \sigma_0^2 = \hat{\sigma}^2$
- This is a special case of the g-prior.

Zellner's g-Prior

Consider priors of the form

$$\beta \mid \sigma^2 \sim N(b_0, g\sigma^2(X^T X)^{-1})$$
$$1/\sigma^2 \sim G(\nu_0/2, \nu_0\sigma_0^2/2)$$

Here, g is a positive constant. When $b_0 = 0$,

$$\beta \mid Y, \sigma^2 \sim N(\frac{g}{1+g}\hat{\beta}, \frac{g}{1+g}\sigma^2(X^T X)^{-1})$$
$$1/\sigma^2 \mid Y \sim G((\nu_0 + n)/2, (\nu_0\sigma_0^2 + SSR_g)/2)$$

where $SSR_g = Y^T(I - \frac{g}{1+g}X(X^T X)^{-1}X^T)Y$, and I is a n-dimensional square identity matrix.

Zellner's g-Prior

Benefits of Zellner's g Prior

- Sample using Monte Carlo techniques (no MCMC needed)
- Bayesian estimate of β shrinks OLS estimate by the quantity $g/(1+g)$
- Recommend $g = n$ to represent vague information about β
- Invariant to re-parameterization: e.g., change of measurement: measurement of age can be year or month. Let D to be a full ranked matrix,

$$Y = X\beta + \epsilon = XD(D^{-1}\beta) + \epsilon$$

The induced prior on the new coefficient vector is

$$D^{-1}\beta \sim \mathsf{N}(0, g\sigma^2 D^{-1}(X^TX)^{-1}D^{-T}) = \mathsf{N}(0, g\sigma^2([XD]^T[XD])^{-1})$$

Independent Prior on β_j

Previously, we let $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ have a multivariate normal prior. When will it be appropriate to use iid prior on $\beta_j, j = 0, \ldots, p$?

- Unit of measurement of all predictors X_j be the same
- Pre-processing step:
  - Center Y and all predictors $X_1, \ldots, X_p$ to mean zero
  - Scale Y and all predictors $X_1, \ldots, X_p$ to variance one

Independent Normal priors

$$\beta_j \mid \sigma^2 \overset{\text{iid}}{\sim} \mathsf{N}(0, \eta\sigma^2)$$

This is equivalent to

$$\beta \sim \mathsf{N}(0, \Sigma_0), \quad \Sigma_0 = \eta\sigma^2 I_n$$

Independent Normal Prior

Conjugate prior

$$\beta_j \mid \sigma^2 \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \eta\sigma^2)$$
$$1/\sigma^2 \sim \mathsf{Gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$$

$$\beta \mid \sigma^2, Y \sim \mathsf{N}\left((I_n/\eta + X^T X)^{-1} X^T Y, \sigma^2 (I_n/\eta + X^T X)^{-1}\right)$$
$$1/\sigma^2 \mid Y \sim \mathsf{Gamma}((\nu_0 + n)/2, \cdots)$$

Independent heavy-tailed prior

Special case: orthogonal design $X^T X = I$, the MLE $\hat{\beta} = X^T Y$, and

$$\beta_j \mid \sigma^2, Y \sim \mathsf{N}\left(\frac{\eta}{1+\eta}\hat{\beta}_j, \frac{\eta}{1+\eta}\sigma^2\right)$$

For any fixed n and η, when $\hat{\beta}_j$ is very large, probably the true value of β_j is very large, then the shrinkage $E(\beta_j \mid \sigma^2, Y) - \hat{\beta}_j = \frac{1}{1+\eta}\hat{\beta}_j$ is large.

To resolve this un-desirable shrinkage, use heavy-tailed prior, e.g., independent Student t distribution.

$$\beta_j \mid \sigma^2 \overset{\text{iid}}{\sim} t(m, 0, \sqrt{\eta\sigma^2})$$

Hierarchical Representation of Student t prior

$$
\begin{aligned}
&\beta_j \mid \sigma^2 \overset{\text{iid}}{\sim} t(m, 0, \sqrt{\eta\sigma^2}) \\
\iff\ & p(\beta_j \mid \sigma^2) \propto \frac{1}{\sqrt{\eta\sigma^2}} \left[1 + \frac{1}{m}\left(\frac{\beta_j^2}{\eta\sigma^2}\right)\right]^{-\frac{m+1}{2}} \\
\iff\ & \begin{cases} \beta_j \mid \lambda_j & \sim \mathsf{N}(0, \lambda_j) \\ \lambda_j \mid \sigma^2 & \sim \mathsf{IG}(\frac{m}{2}, \frac{m\eta\sigma^2}{2}) \end{cases}
\end{aligned}
$$

Full conditionals of $\beta_0, \ldots, \beta_p, \lambda_0, \ldots, \lambda_p, \sigma^2$ available.

Default value $m = 1$: independent Cauchy prior.
Notice that Cauchy mean does not exist.