Computational
Social Science

Linear Regression Fundamentals

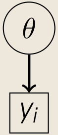
Roberto Cerina

06.02.2024

University of Amsterdam

The Problem of Statistical Inference

1. Observe data from a sample of n units (e.g. individuals):
   $y_i \in \{1, \ldots, n\}$
2. Posit a theory (model) for how the data was generate (Data Generating Process – DGP):
   $y_i \sim f^{\star}$
3. Describe the DGP in terms of some well defined *model parameters*:
   $f^{\star} = f(\theta)$



Note: This representation is called a ‘Directed Acyclic Graph’ (DAG)

The Problem of Statistical Inference

❹ Estimate value of unknown parameter $\theta \rightarrow$ choose the 'most compatible with observed data' (Maximum Likelihood Estimate):
$\hat{\theta}_{MLE} = \arg\max_{\theta} \mathcal{L}(\boldsymbol{y} \mid \theta)$

❺ From the *MLE* we can derive the Empirical Posterior Distribution of θ
$\theta \sim g(\hat{\theta}_{MLE})$

❻ From this distirbution, we can sample *plausible* values of the parameter θ, and make statements about its nature, accounting for uncertainty:
$\theta_s \sim g(\hat{\theta}_{MLE})$ – draw S plausible values of θ;
$\hat{\Pr}(\theta > 0) = \frac{1}{n} \sum_s^n (\theta_s > 0)$ – count how many are > 0 to see if 'significant' (for example ...)

The ‘Homogeneous Expectations’ Gaussian Model

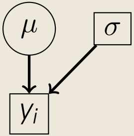
$$y_i \sim N(\mu, \sigma^2) \qquad \forall i \in \{1, ..., n\}$$

- Can also write as: $y_i = \mu + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$.

- μ: expected value (shared across subjects);
- σ: standard deviation from expected value (sometimes noted as σ^2, the ‘variance’, or $\frac{1}{\sigma^2}$ the ‘precision’, also shared across subjects);
- assume σ is known – our interest lies in learning about μ.

- σ^2 represents the ‘unsystematic’ variance in y, that which is inherently random and not predictable;
- $Var(\mu)$ represents the ‘systematic’ variance in y, that which can be understood and systematically predicted;
- % Explained variance: $R^2 = \frac{\text{Var}(\mu)}{\text{Var}(\mu)+\sigma^2}$.

The ‘Homogeneous Expectations’ Gaussian Model



Note:

- ‘square’ nodes indicate ‘observed / known values’;
- ‘circular’ nodes indicate ‘unknown parameters’;
- ‘solid’ arrows indicate ‘stochastic’ relationships – i.e. subject to random variability;
- ‘dotted’ arrows indicate ‘deterministic’ relationships – i.e. a given input will always provide the same output;

Estimation (Point Estimate)

Define a loss function, and minimise !

- Likelihood: $\mathcal{L}_i(\boldsymbol{y} \mid \mu, \sigma = \sigma^\star)$ (sigma is known hence set to $\sigma^\star$)
- for an observation i: $\mathcal{L}_i = \frac{1}{\sqrt{2\pi\sigma^{\star 2}}} \exp\left\{-\frac{1}{2\sigma^{\star 2}}(y_i - \mu)^2\right\}$
- for the entire sample: $\boldsymbol{\mathcal{L}} = \Pi_i^n \mathcal{L}_i$
- log-likelihood: $\log(\boldsymbol{\mathcal{L}}) = \log(\Pi_i^n \mathcal{L}_i) = \sum_i^n \log(\mathcal{L}_i) \leftarrow$ this we want to maximise
- negative log-likelihood: $L = -\sum_i^n \log(\mathcal{L}_i) \leftarrow$ this we want to minimise

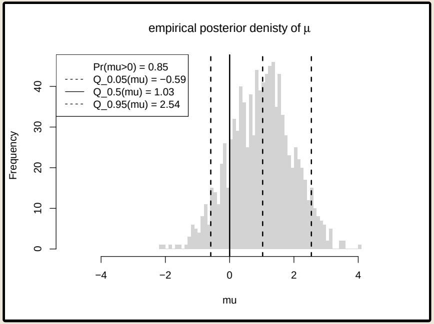
✏ This has an analytical solution:
$\hat{\mu}_{MLE} = \frac{1}{n}\sum_i^n y_i = \bar{y} \leftarrow$ the sample mean is the MLE of μ.

Uncertainty

- MLE has wonderful properties !
- by asymptotic normality ($n \to \infty$) / Central Limit Theorem / Bayesian Posterior Distribution:
  $\mu \sim N(\hat{\mu}_{MLE}, \frac{\sigma^{2\star}}{n})$
- call this the *empirical posterior distribution* of μ.
- *posterior* indicates this is the distribution we learn *after* observing the data.

Simulation-Based (Monte-Carlo) Inference

- What statements can be made about μ ?
- draw S values from the empirical posterior: $\mu_s \sim N(\hat{\mu}_{MLE}, \frac{\sigma^{2\star}}{n})$; $\mu_{1:S} = \{1.85, 1.86, 1.45, 1.29, 1.12, 2.76, 1.69, ...\}$;
- Plot the distribution of μ and infer its properties:
  - statistical significance: what % of the simulated values of μ are larger than 0 ?



Simulation-Based (Monte-Carlo) Inference

‘Monte Carlo’ (MC) methods

- ✎ Calculating statistics about parameters from simulated distributions; Examples:
- ➾ MC Mean (average value of μ across simulations): $\frac{1}{S}\sum_{s}^{S}\mu_{s}$;
- ➾ MC Quantiles (0.05,0.5,0.95) $Q_{\alpha}(\mu_{1:S})$
- ✏ Quantiles are used to get the credibility interval – $Q_{0.5}(\mu_{1:S})$ represents the median, whilst the other quantiles are the lower and upper estimates.

Simulation-Based (Monte-Carlo) Inference

⋆ Note:

- using the simulations method outlined above, we do not refer to 'confidence intervals'...
- the idea of 'confidence' belongs to the realm of hypothesis testing and so called 'frequentist' statistics.
- If you use the empirical posterior to make inference, as we do above, we call these 'credibility intervals';
- these reflect directly the distirbution of plausible or credible values of μ.

Prediction

- Prediction: given what we have learned about our parameters, and the uncertainty associated with this learning, what is our best guess for a new, unseen value of y ?

☞ Prediction too is solved by simulating from the empirical posterior of our parameters !

Prediction

Follow the DGP:

❶ simulate S ‘new’ values from μ according to its empirical posterior distribution:
$\mu_S \sim N(\hat{\mu}_{MLE}, \frac{\sigma^{2\star}}{n})$;

❷ simulate S new values $\tilde{y}$ according to its likelihood, conditional on the simulated values of μ:
$\tilde{y}_S \sim N(\mu_S, \sigma^{2\star})$;

- We call the posterior distribution of $\tilde{y}$, conditional on the posterior distribution of the other model parameters, the *posterior predictive distribution*
- You can then use Monte Carlo methods to make inference about predictions – the MC median will typically be ‘your best guess’.

The ‘Heterogeneous Expectations’ (Multivariate) Gaussian Model

$$
\begin{aligned}
y_i \sim& N(\mu_i, \sigma^2) \qquad \forall i \in \{1, \ldots, n\} \\
\mu_i =& \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}
\end{aligned}
$$

- Can also write as:
  $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon, \qquad \epsilon_i \sim N(0, \sigma^2)$
- In matrix form:
  $\boldsymbol{y} = \beta \boldsymbol{X} + \boldsymbol{e}, \qquad \boldsymbol{e} \sim \text{MN}(\mathbf{0}, \sigma^2 \boldsymbol{I})$
- $\boldsymbol{X}$ is known as the ‘Design Matrix’:

$$
\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}
$$

i. μ_i is the subject-specific expected value:

ii. β_0 is the 'baseline' level of error:
$\rightarrow$ if all other covariates were set to 0 ($\boldsymbol{X} = 0$), the expected level of y;

iii. β_1 ... β_p represent the relationships between variables x_1 ... x_p and the outcome y:
$\rightarrow$ for a 1 unit change in x, we expect to see a change β in y ('controlling for' the effects of the other variables).

The ‘Heterogeneous Expectations’ (Multivariate) Gaussian Model

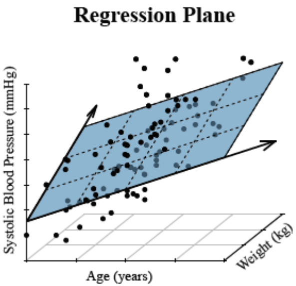


Figure 2.25: Systolic blood pressure linearly increases with age, but also with bodyweight. A line in two directions forms a plane.

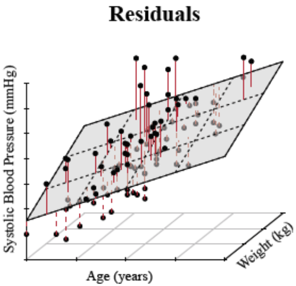
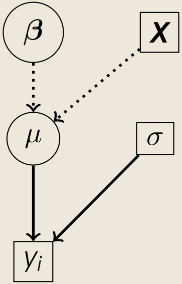


Figure 2.26: The residuals of figure 2.25 are the vertical distances to the plane. Negative residuals are indicated by dashed linepieces.

---

[0] `https://stackoverflow.com/questions/47344850/scatterplot3d-regression-plane-with-residuals`

The ‘Heterogeneous Expectations’ (Multivariate) Gaussian Model



Note:

- ‘square’ nodes indicate ‘observed / known values’;
- ‘circular’ nodes indicate ‘unknown parameters’;
- ‘solid’ arrows indicate ‘stochastic’ relationships – i.e. subject to random variability;
- ‘dotted’ arrows indicate ‘deterministic’ relationships – i.e. a given input will always provide the same output;

Estimation (Point Estimate)

Define a loss function, and minimise !

- Likelihood: $\mathcal{L}_i(\boldsymbol{y} \mid \beta_0, ..., \beta_p, \sigma = \sigma^\star)$ (sigma is known hence set to $\sigma^\star$)
- for an observation i: $\mathcal{L}_i = \frac{1}{\sqrt{2\pi\sigma^{\star 2}}} \exp\left\{-\frac{1}{2\sigma^{\star 2}}(y_i - [\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}])^2\right\}$;
- minimise: $L = -\sum_i^n \log(\mathcal{L}_i)$;

✏ This has an analytical solution by solving a system of equations:
$\hat{\beta}_{MLE} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

Uncertainty

- similar to the 'homogeneous' case, but now we have mmultiple coefficients, and these tend to be correlated ...
- we need a Multivariate Normal distribution to describe the uncertainty around the MLE of $\boldsymbol{\beta}$.

- by asymptotic normality ($n \to \infty$) / Central Limit Theorem / Bayesian Posterior Distribution:

  $\boldsymbol{\beta} \sim MN(\hat{\boldsymbol{\beta}}_{MLE}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^{2\star})$

- $\Sigma = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^{2\star}$ is called the 'Covariance Matrix' of $\boldsymbol{\beta}$.

Simulation-Based (Monte-Carlo) Inference

- What statements can be made about each β_j ?
- draw S simulated values from the *marginal distribution* of β_j:

$$\beta_j^s \sim N(\hat{\beta}_j, \hat{\sigma}^2_{\beta_j}) \qquad \forall s \in \{1, ..., S\}$$

$\hat{\sigma}^2_{\beta_j}$ is simply the j^{th} diagonal element of Σ.

✏ Monte Carlo estimates from these simulated values can reveal significance, and intervals, much like in the univariate case.

Note:

✏ The marginal distribution (which is univariate normal) is sufficient to make inference about a single coefficient amongst those in $\boldsymbol{\beta}$.

✏ If we wanted to make contemporaneous inference about the value of every beta, the *joint distribution* (the multivariate normal) would be necessary.

Prediction

Follow the DGP:

❶ define a set of L ‘new subjects‘ characterised by design vector $\tilde{x}_l, \forall l \in \{1, ..., L\}$;

❷ simulate S ‘new’ values from $\boldsymbol{\beta}$ according to its empirical (joint) posterior distribution:
$\boldsymbol{\beta}^s \sim N(\hat{\boldsymbol{\beta}}_{MLE}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^{2\star})$;

❸ Calculate μ_i for each simulation round:
$\tilde{\mu}_l^s = \beta_0^s + \beta_1^s\tilde{x}_{l1} + ... + +\beta_1^s\tilde{x}_{l1}$

❹ simulate S new values $\tilde{y}$ according to its likelihood, conditional on the simulated values of $\tilde{\mu}$:
$\tilde{y}_l^s \sim N(\tilde{\mu}_l^s, \sigma^{2\star})$;

- You can then use Monte Carlo methods to make inference about predictions – the MC median will typically be ‘your best guess’.