Applied Qual Study Guide

Ting Chen

July 29, 2025

1 Statistical Models

For these models, the following questions are to be answered:

- Model assumptions
- Estimation. Usually there are more than one way to estimate model parameters, each of which arises from their own context and requires different assumptions
- Inference questions: Frequentist distribution, confidence intervals, posterior-distribution based uncertainty measures, etc.
- Model diagnosis and refinement; robustness of estimation and inference to assumptions.
- Model selection/regularization and their computation

1.1 Linear model

BLUE

- Best (least variance)
- Linear
- Unbiased
- Estimator

Gauss-Markov Theorem - no better linear unbiased estimator exists.

Proof:
Consider linear estimate of $\hat{\beta} = \sum_{i=1}^{n} a_i(y_i - \bar{y})$. Then the bias is

$$\mathbb{E}_\varepsilon[\hat{\beta}] = \mathbb{E}_\varepsilon\left[\sum_{i=1}^{n} a_i(\alpha + \beta x_i + \varepsilon_i - \bar{y})\right] = \mathbb{E}_\varepsilon\left[\sum_{i=1}^{n} a_i(\bar{y} - \beta\bar{x} + \beta x_i + \varepsilon_i - \bar{y})\right] = \beta\sum_{i=1}^{n} a_i(x_i - \bar{x})$$

and the variance is

$$\begin{aligned}
\texttt{Var}_\varepsilon[\hat{\beta}] &= \texttt{Var}_\varepsilon[\hat{\beta} - \beta] \\
&= \texttt{Var}_\varepsilon\left[\sum_{i=1}^{n} a_i(y_i - \bar{y}) - \beta\right] \\
&= \texttt{Var}_\varepsilon\left[\sum_{i=1}^{n} a_i(\beta(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})) - \beta\right] \\
&= \texttt{Var}_\varepsilon\left[\beta\sum_{i=1}^{n} a_i(x_i - \bar{x}) + \sum_{i=1}^{n} a_i(\varepsilon_i - \bar{\varepsilon}) - \beta\right] \\
&= \texttt{Var}_\varepsilon\left[\sum_{i=1}^{n} a_i(\varepsilon_i - \bar{\varepsilon})\right] \\
&= \texttt{Var}_\varepsilon\left[\sum_{i=1}^{n} \varepsilon_i(a_i - \bar{a})\right] \\
&= \sigma_\varepsilon^2 \sum_{i=1}^{n} (a_i - \bar{a})^2
\end{aligned}$$

To show the OLS estimates are BLUE, we then solve the constrained minimization problem via Lagrangian multipliers.

$$\begin{aligned}
\min_{a_1,\dots,a_n} \quad & \sum_{i=1}^{n}(a_i - \bar{a})^2 = \sum_{i=1}^{n} a_i^2 - n\bar{a} \\
\text{s.t.} \quad & \sum_{i=1}^{n} a_i(x_i - \bar{x}) = 1
\end{aligned}$$

Taking the derivative wrt to a_i and plugging back into the constraint to get a value for λ yields

$$a_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

as desired.

1.1.1 Model assumptions

1. Gaussian errors - not really needed
2. Homoskedasdicity
3. Additive and linear relationship
4. errors are i.i.d. - not really needed, just uncorrelated and homoskedastic errors
5. zero mean errors

When x and y are standardized, the regression line always has slope less than 1. Thus, when x is 1 standard deviation above the mean, the predicted value of y is somewhere between 0 and 1 standard deviations above the mean. This phenomenon in linear models—that y is predicted to be closer to the mean (in standard-deviation units) than x—is called regression to the mean and occurs in many vivid contexts.

1.1.2 Estimation

1. (O)Least Squares, directly, maximum likelihood estimate:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

for $i = 1, \dots, n$. Want to minimize SSE

$$SSE(\alpha, \beta) = \sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2$$

Taking the derivatives and solving, we get

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\texttt{cov}(x, y)}{\texttt{var}(x)} = \rho_{x,y} \cdot \frac{s_y}{s_x}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Where $s_y = \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, $s_x = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (Note: This form of α implies that the regression line must pass through ($\bar{x}$, $\bar{y}$)), and

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{s_x s_y} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

You get regression to the mean if $\rho_{x,y} < 1$. Some useful properties include

(a) $\sum_{i=1}^{n} \hat{\epsilon}_i = 0 \leftarrow$ take derivative of SSE wrt α

(b) $\sum_{i=1}^{n} x_i\hat{\epsilon}_i = 0 \leftarrow$ take derivative of SSE wrt β

(c) $\sum_{i=1}^{n} \hat{y}_i\hat{\epsilon}_i = 0 \leftarrow$ consequence of the above

which is a consequence of the first order conditions.

Note

$$SSE = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] + \mathbb{E}[(\mathbb{E}(Y|X) - (a + bX))^2]$$

(Cross term drops because noise is independent), hence least squares estimate is best linear approximation to $\mathbb{E}[Y|X = x]$.

Thought experiment assuming X is standard Gaussian, can show via Stein's identity that by minimizing MSE, we are estimating slope of regression function (averaged derivative under Gaussian).

Also note that the error variance is

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_i r_i^2$$

where $r_i := y_i - \hat{y}_i = y_i - (\hat{\alpha} - \hat{\beta}x_i)$.

2. Gradient descent/Newton-Raphson if more params than observations or multicollinearity, can go for regularization to solve this too,

3. Moore-Penrose pseudo-inverse,

4. Bayesian methods (MAP, MCMC, VI, etc.)

1.1.3 Inference questions

The sampling distribution of the estimates slope, intercept and residual variance, conditional on $x_1, \ldots, x_n$, are as follows: $\bar{\epsilon} = 0$ from distribution of errors,

$$\hat{\beta} = \beta + \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Since $\bar{y} \perp \hat{\beta}\bar{x}$,

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \sim \mathcal{N}\left(\alpha, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]\right)$$

Finally

$$\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-2}/(n-2)$$

and note that $(\hat{\alpha}, \hat{\beta}) \perp \hat{\sigma}^2$. Consider the linear regression model:

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N_n(0, \sigma_0^2 I).$$

The least squares estimator is:

$$\hat{Y} = HY, \quad \text{where } H = X(X^\top X)^{-1}X^\top.$$

Then the residual vector is:

$$e = Y - \hat{Y} = (I - H)Y = (I - H)\varepsilon,$$

because $HX\beta = X\beta$.
The residual sum of squares (RSS) is:

$$\text{RSS} = e^\top e = \varepsilon^\top (I - H)\varepsilon.$$

Now apply the **idempotent matrix chi-square theorem** see link here:

- $\varepsilon \sim N_n(0, \sigma_0^2 I)$
- $I - H$ is symmetric and idempotent
- $\text{rank}(I - H) = n - \text{rank}(H) = n - p$, where $p =$ number of parameters in β

In simple linear regression, $p = 2$, so:

$$\frac{1}{\sigma_0^2}\varepsilon^\top (I - H)\varepsilon \sim \chi^2_{n-2}.$$

Hence,

$$\hat{\sigma}^2 = \frac{1}{n-2}\varepsilon^\top (I - H)\varepsilon \sim \frac{\sigma_0^2}{n-2}\chi^2_{n-2}.$$

Confidence intervals on coefficients with t-dist, Compare models with F-test, Test variance with χ^2-test.

Interpret coefficients: "Also the coefficient on sex is more interpretable as it directly represents on average, keeping all other independent variables constant, the average increase/decrease in the tests scores of men compared to women."

1.1.4 Model diagnosis and refinement

- Autocorrelation
- multicollinearity - use instrumental variables
- Linearity and additivity violated, use log transformation -We prefer natural logs (that is, logarithms base e) because, as described above, coefficients on the natural-log scale are directly interpretable as approximate proportional differences
- correlated errors or latent variables to capture violations of the independence assumption, and models for varying variances and nonnormal errors.
- Using observed data to represent a larger population, Duplicate observations, Unequal variances - Weighted regression
- Leverage - point furthest away from $\bar{x}$ has most leverage

1.1.5 Model selection/regularization

L1/L2 regularization, use cross validation/valdiation set for model selection, Adjusted-R^2

1.2 Logistic regression

1.2.1 Model assumptions

1.2.2 Estimation

1.2.3 Inference questions

1.2.4 Model diagnosis and refinement

1.2.5 Model selection/regularization

1.3 Non-parametric models

1.3.1 Model assumptions

1.3.2 Estimation

1.3.3 Inference questions

1.3.4 Model diagnosis and refinement

1.3.5 Model selection/regularization

1.4 Models with latent components including mixed-effect/multilevel models, factor models, etc.

1.4.1 Model assumptions

1.4.2 Estimation

1.4.3 Inference questions

1.4.4 Model diagnosis and refinement

1.4.5 Model selection/regularization

2 Bayesian Data Analysis

Applied and computational Bayesian statistics

2.1 Bayesian Hierarchical Modeling

2.2 Fake-data simulation to design an experiment

2.3 Modeling using splines/Gaussian processes

2.4 Computational workflow

3 Statistical Machine Learning

3.1 Linear and nonlinear dimensionality reduction

3.2 Data-driven and model-based classification methods

3.3 Data-driven and model-based clustering methods

3.4 Graphical models: Bayesian networks, Markov random fields

3.5 Latent variable models

3.6 Introduction to Deep Learning: Deep generative models, Approximate inference

4 Computation

4.1 Gradient-based optimization methods

4.2 Monte Carlo methods: sampling from univariate and multivariate distributions