

Applied Qual Study Guide

Ting Chen

August 3, 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

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

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, can be dropped if sample size is large
2. Homoskedasticity
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

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

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

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

Sampling distributions

The sampling distribution of the estimates slope, intercept and residual variance, conditional on x_1, \dots, x_n , are as follows:

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

note to derive the above we use the fact that the sum of deviations from the mean is always zero, i.e. $\sum_{i=1}^n (x_i - \bar{x}) = 0$.

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

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

and note that $(\hat{\alpha}, \hat{\beta}) \perp \hat{\sigma}^2$.

Proof: Distribution of Residual Variance using Idempotent Matrix χ^2 Theorem
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:

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

because $HX\beta = X\beta$.

The residual sum of squares (RSS) is:

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

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:

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

Hence,

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

Confidence intervals on coefficients with t -dist

Under $H_0 : \beta = 0$

$$\frac{\hat{\beta}}{\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2}$$

Under $H_0 : \alpha = 0$

$$\frac{\hat{\alpha}}{\hat{\sigma} \sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}} \sim t_{n-2}$$

ANOVA (analysis of variance)

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

Coefficient of determination:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Note for OLS $R^2 = \rho_{X,Y}^2$

Proof:

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

and

$$\begin{aligned} R^2 &= \frac{SSR}{SST} \\ &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n (\hat{\alpha} + \hat{\beta}x_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{(\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}))^2}{(\sum_{i=1}^n (y_i - \bar{y})^2)(\sum_{i=1}^n (x_i - \bar{x})^2)} \end{aligned}$$

Compare models with F -test

Measure goodness of fit of your model. Using facts that $SSE \perp SSR$, $SSE \sim \sigma^2 \chi_{n-2}^2$, $SSR \sim \sigma^2 \chi_1^2$ then F -test for $H_0 : \beta = 0$ is

$$F = \frac{SSR}{SSE/(n-2)} \sim F_{1,n-2}$$

Note that the p -value for the F -test and t -test for β are equal in the simple linear regression case.

Prediction intervals

For new data x_{new} , our estimate $\hat{y}_{\text{new}} = \hat{\alpha} + x_{\text{new}}\hat{\beta}$ is unbiased. The variance is

$$\begin{aligned} \text{Var}(\hat{y}_{\text{new}} | x, x_{\text{new}}) &= \text{Var}(\hat{\alpha} | x) + x_{\text{new}}^2 \text{Var}(\hat{\beta} | x, x_{\text{new}}) + 2x_{\text{new}} \text{Cov}(\hat{\alpha}, \hat{\beta}) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \end{aligned}$$

where $\text{Cov}(\hat{\alpha}, \hat{\beta}) = \frac{-\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$. For proof of this consider the following:

$$\begin{aligned}
\text{Cov}(\hat{\alpha}, \hat{\beta}) &= \text{Cov}(\bar{y} - \hat{\beta}\bar{x}, \hat{\beta}) \\
&= \text{Cov}(\bar{y}, \hat{\beta}) - \text{Cov}(\hat{\beta}\bar{x}, \hat{\beta}) \\
&= \text{Cov}\left(\frac{1}{n} \sum_{i=1}^n y_i, \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2}\right) - \bar{x} \text{Var}(\hat{\beta}) \\
&= \frac{\sum_{i=1}^n \sigma^2 (x_i - \bar{x})}{n \sum_{j=1}^n (x_j - \bar{x})^2} - \frac{\sigma^2 \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2} \quad (\text{See Lemma 11.3.2. from Casella and Berger}) \\
&= 0 - \frac{\sigma^2 \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2}
\end{aligned}$$

Hence

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

and it follows that a CI to use would be

$$\hat{\alpha} + x_{\text{new}} \cdot \hat{\beta} \pm t_{n-2, 1-\alpha/2} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

However we're typically interested in an interval for the actual observations rather than on the mean. Hence

$$\begin{aligned}
\text{Var}(y_{\text{new}} - \hat{y}_{\text{new}}) &= \text{Var}(y_{\text{new}}) + \text{Var}(\hat{y}_{\text{new}}) \\
&= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)
\end{aligned}$$

hence the CI we do use is

$$\hat{\alpha} + x_{\text{new}} \cdot \hat{\beta} \pm t_{n-2, 1-\alpha/2} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Relaxing assumptions and their impacts on CIs:

1. Normality

- Check with Q-Q plot of residuals
- Can be dropped with large sample sizes as by (Lindeberg-Feller) CLT note that

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

and

$$\hat{\alpha} \xrightarrow{d} \mathcal{N}\left(\alpha, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right)$$

However in this regime $\hat{\alpha}$ and $\hat{\beta}$ are **not independent** of $\hat{\sigma}^2$ and hence we must use Slutsky's to justify using normal quantiles in our confidence intervals (the side effect here is also that the use of t -distribution quantiles no longer become valid).

2. Linearity

- Check with residual vs. fitted value plots
- When there is nonlinearity and $\alpha + \beta X$ are still the best linear approximation, then point estimates and standard errors are still valid but the interpretations are different (this is just the best linear approximation). Consider

$$\mathbb{E}[Y|X] = \alpha + \beta X + \delta(X)$$

If $\alpha + \beta X$ is the best linear approximation then (assuming X is random)

$$\mathbb{E}[\delta(X)] = \mathbb{E}[X\delta(X)] = 0$$

(α is best intercept and β is the best linear term). In which case we still have

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

and

$$\hat{\alpha} \xrightarrow{d} \mathcal{N}\left(\alpha, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right)$$

where $\sigma^2 = \sigma_0^2 + \mathbb{E}[\delta(X)^2]$ Finally, note that

$$\hat{\sigma}^2 \xrightarrow{p} \sigma^2 = \sigma_0^2 + \mathbb{E}[\delta(X)^2] > \sigma_0^2$$

3. Homoskedasticity

- Check with residual vs. fitted value plots
- If we drop this, our point estimates remain valid, but the standard errors and inferences need to be adjusted. Consider $\text{Var}(\varepsilon_i) = \sigma_i^2$, then

$$\text{Var}(\hat{\beta}) = \frac{\sum_{i=1}^n \sigma_i^2 (x_i - \bar{x})^2}{(\sum_{j=1}^n (x_j - \bar{x})^2)^2}$$

Since we can't directly estimate σ_i^2 , we use the following, justified by Slutsky's

$$\widehat{\text{Var}}(\hat{\beta}) := \frac{\sum_{i=1}^n r_i^2 (x_i - \bar{x})^2}{(\sum_{j=1}^n (x_j - \bar{x})^2)^2} \xrightarrow{p} \text{Var}(\hat{\beta})$$

4. Independence of residuals

- Check with residual vs. fitted value plots
- When $\text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma_{ij}$, then

$$\text{Var}(\hat{\beta}) = \frac{\sum_{i,j} \sigma_{ij} (x_i - \bar{x})(x_j - \bar{x})}{(\sum_i (x_i - \bar{x})^2)^2}$$

The CLT still holds under weak dependence (triangular CLT).

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

and

$$\hat{\alpha} \xrightarrow{d} \mathcal{N}\left(\alpha, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right)$$

Point estimates are still valid, standard errors may be valid.

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/validation set for model selection, Adjusted- R^2

1.1.6 Notes from past problems

- Applied Qual 2024 Problem 2
 - (a) Fitting a single regression line with a binary predictor between two groups and interaction term with the continuous predictor is equivalent to fitting two separate regression lines to the two groups since the degrees of freedom are the same and we assume noise is independent so fitting of one line will not affect the other.
 - (b)

$$\hat{\beta} = \frac{\text{cov}(x, y)}{\text{var}(x)} = \rho_{x,y} \cdot \frac{s_y}{s_x}$$

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

- (c) Simpson's paradox
- (d) Empirical bootstrap procedure
 - (i) Sample with replacement from data n times
 - (ii) Fit regression model to sampled data
 - (iii) Repeat step i and ii B times to get $\hat{\beta}_2^{(1)}, \dots, \hat{\beta}_2^{(B)}$
 - (iv) By asymptotic theory, we know that there exist σ_j for $j \in [3]$, such that

$$\sqrt{n}(\hat{\beta}_j - \beta_j) \xrightarrow{d} \mathcal{N}(0, \sigma_j^2)$$

Hence we can construct approximate C.I.s of the form

$$\hat{\beta}_j \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\text{Var}}_B(\hat{\beta}_j)}$$

where we estimate $\widehat{\text{Var}}_B(\hat{\beta}_j) = \frac{1}{B} \sum_{i=1}^B \left(\hat{\beta}_j^{(i)} - \frac{1}{B} \sum_{k=1}^B \hat{\beta}_j^{(k)} \right)^2$ from the bootstrap samples.

- (e) Generally $\rho_{x,y} < 1$ (noise with non-zero variance), hence flipping will not yield same estimate.

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