Applied Qual Study Guide

Ting Chen

August 6, 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.

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

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 $i = 1, \ldots, n$. 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

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

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

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:

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

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
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

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

Residual standard error (RSE):

$$RSE = \sqrt{\frac{SSR}{n-p-1}}$$

Compare models with F-test
Measure goodness of fit of your model. Using facts that $SSE \perp SSR$, $SSE \sim \sigma^2\chi^2_{n-2}$, $SSR \sim \sigma^2\chi^2_1$ 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

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

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

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

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

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

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 $\texttt{Var}(\varepsilon_i) = \sigma_i^2$, then
   $$\texttt{Var}(\hat{\beta}) = \frac{\sum_{i=1}^n \sigma_i^2 (x_i - \bar{x})^2}{(\sum_{j=1}^n (x_j - \bar{x})^2)^2}$$
   Since we can't directly estimate σ_i^2, we use the following, justified by Slutsky's
   $$\widehat{\texttt{Var}(\hat{\beta})} := \frac{\sum_{i=1}^n r_i^2 (x_i - \bar{x})^2}{(\sum_{j=1}^n (x_j - \bar{x})^2)^2} \xrightarrow{p} \texttt{Var}(\hat{\beta})$$

4. Independence of residuals
   - Check with residual vs. fitted value plots
   - When $\texttt{Cov}(\varepsilon_i, \varepsilon_j) = \sigma_{ij}$, then
   $$\texttt{Var}(\hat{\beta}) = \frac{\sum_{i,j} \sigma_{ij}(x_i - \bar{x})(x_j - \bar{x})}{(\sum_i (x_i - \bar{x})^2)^2}$$
   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/valdiation 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)
  $$\hat{\beta} = \frac{\texttt{cov}(x, y)}{\texttt{var}(x)} = \rho_{x,y} \cdot \frac{s_y}{s_x}$$
  $$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$
  (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 $\hat{\beta}_2^{(1)}, \ldots, \hat{\beta}_2^{(B)}$
    (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
    $$\hat{\beta}_j \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\texttt{Var}_B(\hat{\beta}_j)}}$$
    where we estimate $\widehat{\texttt{Var}_B(\hat{\beta}_j)} = \frac{1}{B}\sum_{i=1}^{B}\left(\hat{\beta}_j^{(i)} - \frac{1}{B}\sum_{k=1}^{B}\hat{\beta}_j^{(k)}\right)^2$ from the bootstrap samples.

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

- Applied Qual 2022 Problem A

  (a) Ablate rounding to nearest 3 months and no rounding. Then can compare following models:

  - Standard OLS - issues include heteroskedastic noise (heights will vary more with age), fact that ages can't be less than 0, and would expect growth spurts so additive linear assumption is not correct.
$$y_i|x_i \sim \text{Normal}(\alpha + \beta x_i, \sigma^2)$$
  - Log transformation of y's (heights) remedies the second two issues from above a bit.
$$y_i|x_i \sim \text{LogNormal}(\alpha + \beta x_i, \sigma^2)$$
  - Another approach could be to use a latent variable/hierarchical model, where we model the latent true age t_i using a categorical latent variable δ_i.
$$y_i|t_i \sim \text{LogNormal}(\alpha + \beta t_i, \sigma^2)$$
$$\delta_i \sim \text{Categorical}(\pi_1, \pi_2, \pi_3)$$

  Where π_j for $j \in [3]$ corresponds to the probability that $\delta_i = j$ and
$$\delta_i = \begin{cases} 1 & \text{if age is exact. Hence } t_i = x_i. \\ 2 & \text{if age is rounded to nearest 6 months. Hence } t_i \in [x_i - 3, x_i + 3). \\ 3 & \text{if age is rounded to nearest 12 months. Hence } t_i \in [x_i - 6, x_i + 6). \end{cases}$$

  For a prior on t_i, we assume that
$$t_i \sim \text{Uniform}(0, 60)$$

  Hence the (global) parameters that we need to estimate are $\theta = \{\alpha, \beta, \pi_1, \pi_2, \pi_3, \sigma^2\}$.

  (b) Using second model, likelihood is just product of LogNormal pdfs
$$\begin{aligned}\mathcal{L}(\alpha, \beta, \sigma^2 | \{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n) &= \prod_{i=1}^n \frac{1}{x_i\sqrt{2\pi\sigma^2}} \exp\left(\frac{1}{2\sigma^2}(\log y_i - \alpha - \beta x_i)^2\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(\sum_{i=1}^n \frac{1}{2\sigma^2}(\log y_i - \alpha - \beta x_i)^2 - \log y_i\right)\end{aligned}$$

Using the third model, the likelihood is

$$\begin{aligned}
\mathcal{L}(\theta|\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n)) &= \prod_{i=1}^n p(y_i, x_i|\theta) \\
&= \prod_{i=1}^n \sum_{j=1}^3 \int_0^{60} p(y_i, x_i, t_i, \delta_i = j|\theta) dt \\
&= \prod_{i=1}^n \sum_{j=1}^3 \int_0^{60} p(y_i, |t_i, \theta) p(x_i, |\delta_i = j, t_i, \theta) p(\delta_i = j|\theta) p(t_i|\theta) dt \\
&= \prod_{i=1}^n \sum_{j=1}^3 \int_0^{60} \text{LogNormal}(\alpha + \beta t_i) \cdot \mathbb{1}(\delta_{ij}(x_i)) \cdot \pi_j \cdot \frac{1}{60} dt \\
&= \prod_{i=1}^n \frac{1}{60} \left(\int_0^{60} \text{LogNormal}(\alpha + \beta t_i) \cdot \mathbb{1}(t_i = x_i) \cdot \pi_1 dt \right. \\
&+ \int_0^{60} \text{LogNormal}(\alpha + \beta t_i) \cdot \mathbb{1}(t_i \in [x_i - 3, x_i + 3)) \cdot \pi_2 dt \\
&+ \left. \int_0^{60} \text{LogNormal}(\alpha + \beta t_i) \cdot \mathbb{1}(t_i \in [x_i - 6, x_i + 6)) \cdot \pi_3 dt \right) \\
&= \prod_{i=1}^n \frac{1}{60} \left(\text{LogNormal}(\alpha + \beta x_i) \cdot \pi_1 + \int_{x_i-3}^{x_i+3} \text{LogNormal}(\alpha + \beta t_i) \cdot \pi_2 dt \right. \\
&+ \left. \int_{x_i-6}^{x_i+6} \text{LogNormal}(\alpha + \beta t_i) \cdot \pi_3 dt \right)
\end{aligned}$$

Note in the second line we marginalize over the (local) latent variables t_i, δ_i.

(c) Using second model, do MLE directly. Note the MLEs for LogNormal regression is equivalent to the MLEs for OLS, except with the y_i's replaced with $\log y_i$'s. Hence

$$\begin{aligned}
\hat{\beta} &= \frac{\texttt{Cov}(x, \log y)}{\texttt{Var}(x)} \\
\hat{\alpha} &= \overline{\log y} - \hat{\beta}\bar{x} \\
\widehat{\sigma^2} &= \frac{1}{n} \sum_{i=1}^n (\log y_i - (\alpha + \beta x_i))^2
\end{aligned}$$

Could do gradient descent as well if there are numerical issues. To do inference you could look at the predictive distribution of $\log y_i$ and invert it to get a point estimate of the true age. Then you could do a "Wild" or residual bootstrap (ref) to get a distribution on the true age given a specific height maybe? (probably not right).

Using the third model, you could theoretically directly maximize the log of the above observed data log likelihood via MLE. However notice that we would then have a log of a sum in addition to having to differentiate under the integral. The resulting expression is highly likely to run into numerical issues if you try to use it with a gradient descent type algorithm. As an alternative, we could do EM/MCMC/VI, for simplicity we'll just describe an EM algorithm for this model.

E step: Compute expectations/responsibilities of latent variables using complete data log likelihood (likelihood of global parameters assuming you have observations for local latent variables).

Also can be seen as estimating the posterior of the local latent variables (MAP estimate). Here the complete data log likelihood is

$$\begin{aligned}
\log \mathcal{L}_C(\theta|\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n, \{t_i\}_{i=1}^n, \{\delta_i\}_{i=1}^n) &= \sum_{i=1}^n \log p(y_i, x_i, t_i, \delta_i|\theta) \\
&= \sum_{i=1}^n \log p(y_i|t_i, \theta) + \log p(x_i|t_i, \delta_i, \theta) + \log p(t_i) + \log p(\delta_i|\theta) \\
&\propto \sum_{i=1}^n \left(-\frac{1}{2}\log \sigma^2 + \frac{1}{2\sigma^2}(\log y_i - \alpha - \beta t_i)^2 + \sum_{j=1}^3 \mathbb{1}(\delta_i = j)\log \pi_j \right)
\end{aligned}$$

where we drop terms that do not depend on θ (i.e. $\log p(x_i|t_i, \delta_i, \theta)$ and $\log p(t_i)$). Then for the E step, given an initial guess for $\theta^{(0)}$, we compute

$$\mathbb{E}_{\mathbf{t},\delta|\mathbf{x},\mathbf{y},\theta^{(0)}}[\log \mathcal{L}_C(\theta|\phi)] \propto \mathbb{E}_{\mathbf{t},\delta|\mathbf{x},\mathbf{y},\theta^{(0)}} \left[\sum_{i=1}^n \left(-\frac{1}{2}\log \sigma^2 + \frac{1}{2\sigma^2}(\log y_i - \alpha - \beta t_i)^2 + \sum_{j=1}^3 \mathbb{1}(\delta_i = j)\log \pi_j \right) \right]$$

where we use ϕ as a shorthand for the complete data $(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n, \{t_i\}_{i=1}^n, \{\delta_i\}_{i=1}^n)$. Inspecting the above expression, we notice we need to compute three expressions:

(1)

$$\begin{aligned}
\mathbb{E}_{\mathbf{t},\delta|\mathbf{x},\mathbf{y},\theta^{(0)}}[\mathbb{1}(\delta_i = j)] &= \sum_{\delta=1}^3 \int \mathbb{1}(\delta_i = j) p(\mathbf{t}, \delta|\mathbf{x}, \mathbf{y}, \theta^{(0)}) d\mathbf{t} \\
&= \sum_{\delta=1}^3 \int \mathbb{1}(\delta_i = j) p(\mathbf{t}|\delta, \mathbf{x}, \mathbf{y}, \theta^{(0)}) p(\delta|\mathbf{x}, \mathbf{y}, \theta^{(0)}) d\mathbf{t} \\
&= p(\delta_i = j|\mathbf{x}, \mathbf{y}, \theta^{(0)}) \int p(\mathbf{t}|\delta, \mathbf{x}, \mathbf{y}, \theta^{(0)}) d\mathbf{t} \\
&= p(\delta_i = j|\mathbf{x}, \mathbf{y}, \theta^{(0)})
\end{aligned}$$

where above we use the chain rule. To actually compute this posterior probability, we appeal to Bayes rule:

$$\begin{aligned}
p(\delta_i = j|\mathbf{x}, \mathbf{y}, \theta^{(0)}) &= \frac{p(\delta_i = j, \mathbf{x}, \mathbf{y}|\theta^{(0)})}{p(\mathbf{x}, \mathbf{y}|\theta^{(0)})} \\
&= \frac{p(\mathbf{x}, \mathbf{y}|\delta_i = j, \theta^{(0)}) p(\delta_i = j|\theta^{(0)})}{\sum_{k=1}^3 p(\mathbf{x}, \mathbf{y}, \delta_i = k|\theta^{(0)})} \\
&= \frac{p(\mathbf{x}, \mathbf{y}|\delta_i = j, \theta^{(0)}) \pi_j}{\sum_{k=1}^3 p(\mathbf{x}, \mathbf{y}|\delta_i = k, \theta^{(0)}) \pi_k}
\end{aligned}$$

where as above in the likelihood part

$$p(\mathbf{x}, \mathbf{y}|\delta_i = 1, \theta^{(0)}) = \int_0^{60} p(\mathbf{x}, \mathbf{y}, \mathbf{t}|\delta_i = 1, \theta^{(0)})d\mathbf{t} = \int_0^{60} p(\mathbf{x}|\mathbf{t}, \delta_i = 1, \theta^{(0)})p(\mathbf{y}|\mathbf{t}, \delta_i = 1, \theta^{(0)})p(\mathbf{t})d\mathbf{t}$$
$$= \text{LogNormal}(\alpha + \beta x_i) \cdot \frac{1}{60}$$

$$p(\mathbf{x}, \mathbf{y}|\delta_i = 2, \theta^{(0)}) = \int_0^{60} p(\mathbf{x}, \mathbf{y}, \mathbf{t}|\delta_i = 1, \theta^{(0)})d\mathbf{t} = \int_0^{60} p(\mathbf{x}|\mathbf{t}, \delta_i = 1, \theta^{(0)})p(\mathbf{y}|\mathbf{t}, \delta_i = 1, \theta^{(0)})p(\mathbf{t})d\mathbf{t}$$
$$= \int_{\mathbf{x}-3}^{\mathbf{x}+3} \text{LogNormal}(\alpha + \beta\mathbf{t}) \cdot \frac{1}{60}d\mathbf{t}$$

$$p(\mathbf{x}, \mathbf{y}|\delta_i = 3, \theta^{(0)}) = \int_0^{60} p(\mathbf{x}, \mathbf{y}, \mathbf{t}|\delta_i = 1, \theta^{(0)})d\mathbf{t} = \int_0^{60} p(\mathbf{x}|\mathbf{t}, \delta_i = 1, \theta^{(0)})p(\mathbf{y}|\mathbf{t}, \delta_i = 1, \theta^{(0)})p(\mathbf{t})d\mathbf{t}$$
$$= \int_{\mathbf{x}-6}^{\mathbf{x}+6} \text{LogNormal}(\alpha + \beta\mathbf{t}) \cdot \frac{1}{60}d\mathbf{t}$$

(2)

$$\mathbb{E}_{\mathbf{t},\delta|\mathbf{x},\mathbf{y},\theta^{(0)}}[t_i] = \sum_{k=1}^{3} \int \mathbf{t} \cdot p(\mathbf{t}, \delta = k|\mathbf{x}, \mathbf{y}, \theta^{(0)})d\mathbf{t}$$
$$= \sum_{k=1}^{3} \int \mathbf{t} \cdot p(\mathbf{t}|\delta = k, \mathbf{x}, \mathbf{y}, \theta^{(0)})p(\delta = k|\mathbf{x}, \mathbf{y}, \theta^{(0)})d\mathbf{t}$$

where again by Bayes rule

$$p(\mathbf{t}|\delta = k, \mathbf{x}, \mathbf{y}, \theta^{(0)}) = \frac{p(\delta = k, \mathbf{x}, \mathbf{y}, \mathbf{t}|\theta^{(0)})}{p(\delta = k, \mathbf{x}, \mathbf{y}|\theta^{(0)})}$$
$$= \frac{p(\delta = k|\theta^{(0)})p(\mathbf{x}, \mathbf{y}, \mathbf{t}|\delta = k, \theta^{(0)})}{p(\delta = k|\theta^{(0)})p(\mathbf{x}, \mathbf{y}|\delta = k, \theta^{(0)})}$$
$$= \frac{\text{LogNormal}(\alpha + \beta t_i) \cdot \mathbb{1}(\delta_{ik}(x_i)) \cdot \frac{1}{60}}{\int_0^{60} p(\mathbf{x}|\mathbf{t}, \delta_i = k, \theta^{(0)})p(\mathbf{y}|\mathbf{t}, \delta_i = k, \theta^{(0)})p(\mathbf{t})d\mathbf{t}}$$

where hence we have already described how to calculate all of the above quantities.

(3) Likewise

$$\mathbb{E}_{\mathbf{t},\delta|\mathbf{x},\mathbf{y},\theta^{(0)}}[t_i^2] = \sum_{k=1}^{3} \int \mathbf{t}^2 \cdot p(\mathbf{t}|\delta = k, \mathbf{x}, \mathbf{y}, \theta^{(0)})p(\delta = k|\mathbf{x}, \mathbf{y}, \theta^{(0)})d\mathbf{t}$$

M step: Maximize expected value of the complete data log likelihood to estimate global parameters θ, analogous to MLE. Update π_k by taking average responsibility over data.

$$Q_\pi(\pi) = \sum_{i=1}^{n} E\left[\sum_{k=1}^{3} I(z_i = k)\log(\pi_k)\right] = \sum_{i=1}^{n}\sum_{k=1}^{3} E[I(z_i = k)]\log(\pi_k)$$

Since $E[I(z_i = k)] = p(z_i = k|a_i, h_i, \theta^{(j)}) = w_{ik}^{(j)}$, this simplifies to:

$$Q_\pi(\pi) = \sum_{i=1}^{n}\sum_{k=1}^{3} w_{ik}^{(j)}\log(\pi_k)$$

We need to maximize this function subject to the constraint that $\sum_{k=1}^3 \pi_k = 1$. We use a Lagrange multiplier, λ.

$$\mathcal{L}(\pi, \lambda) = \sum_{i=1}^n \sum_{k=1}^3 w_{ik}^{(j)} \log(\pi_k) + \lambda(1 - \sum_{k=1}^3 \pi_k)$$

Taking the derivative with respect to π_k and setting it to zero:

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \sum_{i=1}^n \frac{w_{ik}^{(j)}}{\pi_k} - \lambda = 0 \implies \pi_k = \frac{\sum_{i=1}^n w_{ik}^{(j)}}{\lambda}$$

To find λ, we sum over all k and use the constraint:

$$\sum_{k=1}^3 \pi_k = 1 \implies \frac{1}{\lambda} \sum_{k=1}^3 \sum_{i=1}^n w_{ik}^{(j)} = 1$$

The sum $\sum_{k=1}^3 \sum_{i=1}^n w_{ik}^{(j)} = \sum_{i=1}^n \sum_{k=1}^3 w_{ik}^{(j)}$. Since $\sum_{k=1}^3 w_{ik}^{(j)} = 1$ for any child i, the total sum is simply n.

$$\frac{n}{\lambda} = 1 \implies \lambda = n$$

Update α, β, σ^2 by doing a weighted least squares regression, where we use the values for $\mathbb{E}_{\mathbf{t},\delta|\mathbf{x},\mathbf{y},\theta^{(0)}}[t_i]$ and $\mathbb{E}_{\mathbf{t},\delta|\mathbf{x},\mathbf{y},\theta^{(0)}}[t_i^2]$ that we calculated in the E step above in place of t_i and t_i^2.

To do inference on the true age of a child, we calculate

$$p(t_i|\mathbf{x}, \mathbf{y}, \hat{\theta}) = \sum_{k=1}^3 p(t_i|\delta = k, \mathbf{x}, \mathbf{y}, \hat{\theta}) p(\delta = k|\mathbf{x}, \mathbf{y}, \hat{\theta})$$

From this probability distribution we can calculate a point estimate and look at the quantiles if it is a nice distribution or else do an empirical bootstrap.

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