Hoff Ch. 9 : Linear regression

Setup

- Data are pairs (x_i, y_i) where $x_i \in \mathbb{R}^d$ is an observation of a r.v. X_i and $y_i \in \mathbb{R}$ is an observation of a r.v. Y_i.
- $x_i = (x_{i1}, \ldots, x_{id})$ is a vector of <u>covariates</u> or <u>predictors</u>
- y_i is a <u>response</u> or <u>outcome</u>
- If the data $\{(x_i, y_i) : i = 1, \ldots, n\}$ are iid (as is often assumed), then each x_i is an independent observ. of X (so drop the "i") and y_i of Y.

Goal

- We want to model Y given X
  - What is $p(y|x)$?
  - Given newly observed covariates x_{n+1}, predict y_{n+1}
    - * PPD: $p(y_{n+1}|y_1,\dots,y_n)$
    - * Supervised prediction: $p(y_{n+1}|x_{n+1},x_1,\dots,x_n,y_1,\dots,y_n)$
- General (additive) regression model:

$$Y = f(X) + \varepsilon \qquad (*)$$

where $f: \mathbb{R}^d \longrightarrow \mathbb{R}$ is an unknown function and ε is random noise ("additive" refers to the fact that ε is added to $f(X)$).

- This chapter: special case of (*) where $f(X) = \beta^T X = \sum_{j=1}^{d} \beta_j X_j$, i.e., f is a linear function:

$$Y = \beta^T X + \varepsilon.$$

- Typically (though not necessarily), $\varepsilon \sim N(0, \sigma^2)$.

Assumptions/notation throughout these notes

- Assumption 1: Linear model, $Y = \beta^T X + \varepsilon$
- Assumption 2: $\varepsilon \sim N_1(0, \sigma^2)$
- Assumption 3: We'll treat the observed covariates $x_1, \ldots, x_n \in \mathbb{R}^d$ as fixed

- For the observed data, we have

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1d} \\ & & \vdots & \\ X_{n1} & X_{n2} & \cdots & X_{nd} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_d \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or, in more compact notation,

$$\vec{Y}_n = \mathbf{X}\beta + \varepsilon \quad \text{where} \quad \varepsilon \sim N(0, \sigma^2 I_n)$$

$n \times d$ matrix of observed covariates, called the data matrix or design matrix

- In the Bayesian framework, β and σ^2 will be our parameters of interest. However, let's first look at the frequentist approach to linear regression, assuming n (# of samples) is larger than d (# of covariates).

- Under the assumption that $\varepsilon \sim N(0, \sigma^2 I_n)$, we have

$$\vec{Y}_n \mid \mathbf{X}, \beta, \sigma^2 \sim N_n(\mathbf{X}\beta, \sigma^2 I_n).$$

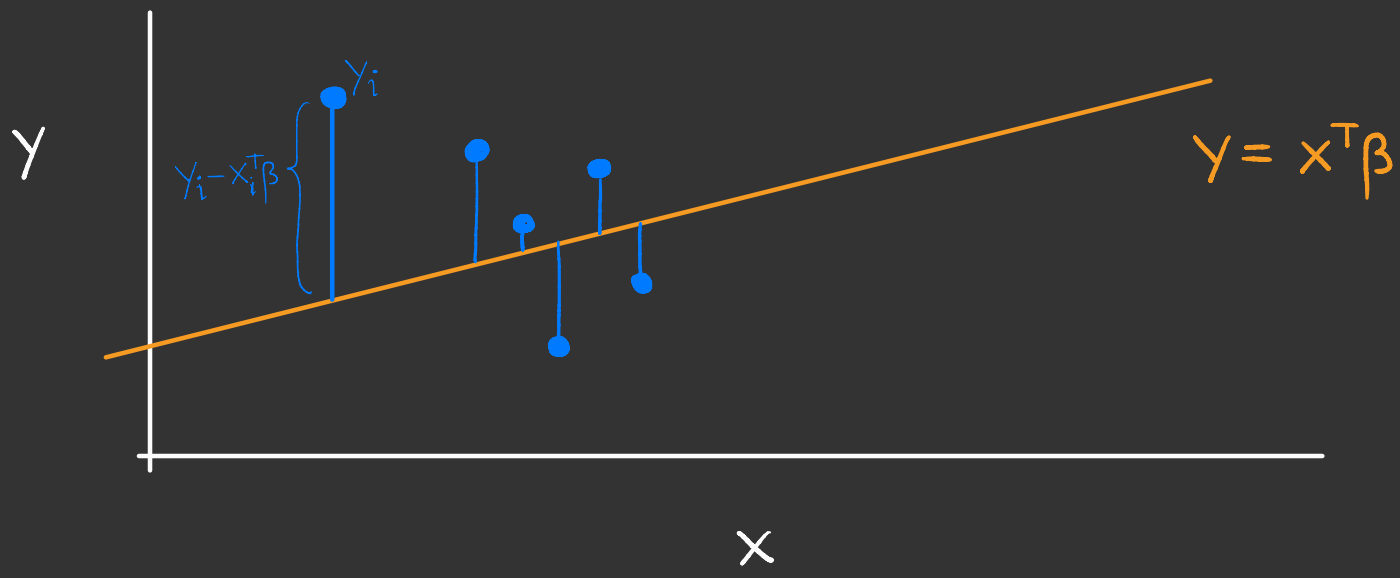
The density of $\vec{Y}_n$ given $\mathbf{X}, \beta, \sigma^2$ is therefore

$$p(\vec{Y}_n \mid \mathbf{X}, \beta, \sigma^2) = \left((2\pi)^n \sigma^{2n}\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(\vec{Y}_n - \mathbf{X}\beta)^T(\vec{Y}_n - \mathbf{X}\beta)\right]$$

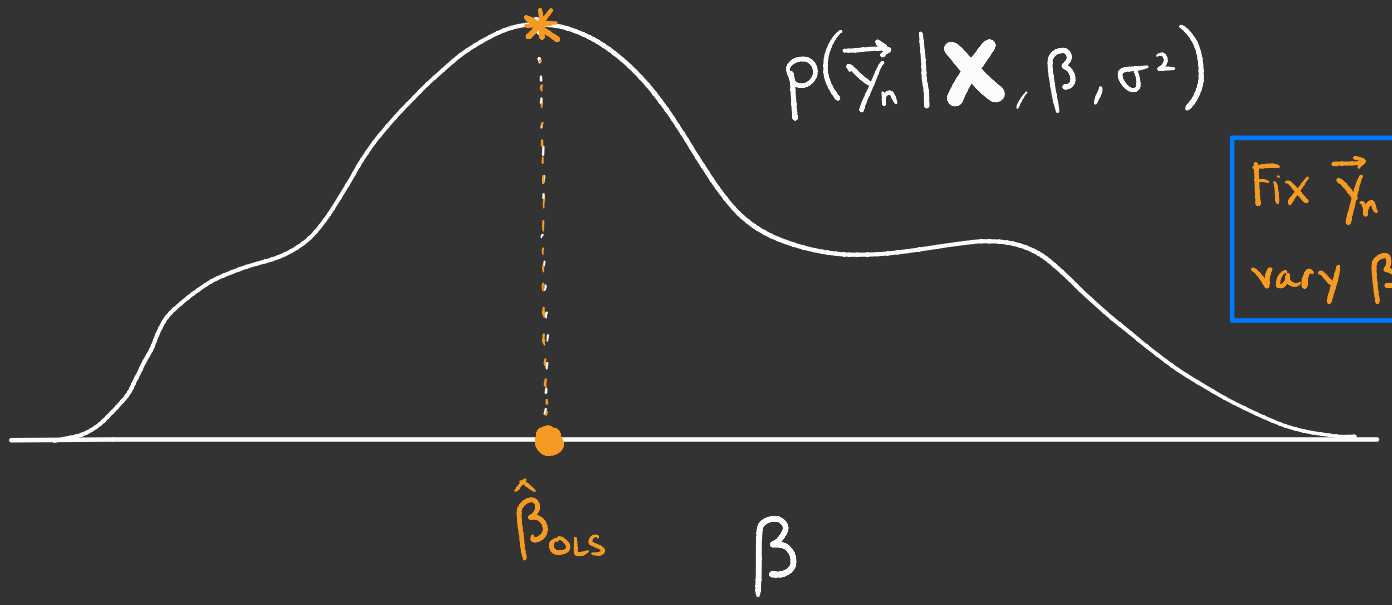
$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - x_i^T\beta)^2\right]$$

- Define the sum of squared residuals (SSR) by

$$SSR(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$



- The <u>maximum likelihood estimate</u> (MLE) for β, denoted $\hat{\beta}_{OLS}$ (OLS = ordinary least squares) in this case, is the value of β that maximizes the density $p(\vec{y}_n | \mathbf{X}, \beta, \sigma^2)$.



That is, $\hat{\beta}_{OLS} = \underset{\beta}{\operatorname{argmax}}\ p(\vec{y}_n \mid \mathbf{X}, \beta, \sigma^2)$ or, equivalently,

$$\hat{\beta}_{OLS} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i^T\beta)^2 = \underset{\beta}{\operatorname{argmin}}\ (\vec{y}_n - \mathbf{X}\beta)^T(\vec{y}_n - \mathbf{X}\beta)$$

Doesn't depend on β

$$= \underset{\beta}{\operatorname{argmin}}\ \vec{y}_n^T\vec{y}_n - \left(\vec{y}_n^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\vec{y}_n\right) + \beta^T\mathbf{X}^T\mathbf{X}\beta$$

$$= \underset{\beta}{\operatorname{argmin}}\ \beta^T\mathbf{X}^T\mathbf{X}\beta - 2\beta^T\mathbf{X}^T\vec{y}_n$$

For the blue part: Note that $c^T = c$ for any scalar, i.e., $c \in \mathbb{R}$.

So $\vec{y}_n^T\mathbf{X}\beta = (\vec{y}_n^T\mathbf{X}\beta)^T = \beta^T\mathbf{X}^T\vec{y}_n$.

- To find $\hat{\beta}_{OLS}$, recall that minima of a function occur where its derivative is 0 and second derivative is >0.
- We want minimum of $f: \mathbb{R}^d \longrightarrow \mathbb{R}$ given by

$$f(\beta) = \beta^T \mathbf{X}^T \mathbf{X} \beta - 2\beta^T \mathbf{X}^T \vec{y}_n$$

$$\Rightarrow 0 = \frac{d}{d\beta} f(\beta) = 2\mathbf{X}^T \mathbf{X} \beta - 2\mathbf{X}^T \vec{y}_n$$

$$\Rightarrow \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \vec{y}_n$$

$$\Rightarrow \beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{y}_n$$

Second derivative,

$$\partial_\beta^2 f(\beta) = 2\mathbf{X}^T\mathbf{X}$$

is positive definite (matrix analogue of > 0), so

$$\hat{\beta}_{OLS} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\vec{Y}_n$$

- Important aside: The above assumes $\mathbf{X}^T\mathbf{X}$ is invertible, which is only possible if $n \geq d$.

Bayesian linear regression

- The MLE approach gives a single estimate $\hat{\beta}_{OLS}$.
- Bayes approach looks at the full distribution of β and σ^2, i.e., $p(\beta, \sigma^2 \mid \vec{Y}_n, \mathbf{X})$.
- Recall for the univariate normal model we had two different Bayesian models. In both cases,

$$Y_1, \ldots, Y_n \mid \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2).$$

That is, the likelihood is the same. What differed

were the priors:

- Bayesian model #1 (semi-conjugate):
  - $p(\theta) = N(\mu_0, \tau_0^2)$
  - $p(\sigma^2) = IG\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right)$

  } independent priors

$\Rightarrow$ full conditionals:

- $p(\theta \mid \sigma^2, \vec{y}_n) = N(\mu_n, \tau_n^2)$
- $p(\sigma^2 \mid \theta, \vec{y}_n) = IG\left(\frac{\nu_n}{2}, \frac{\nu_n}{2}\sigma_n^2\right)$

- Bayesian model #2 (dependent priors):
  - $p(\sigma^2) = IG\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right)$
  - $p(\theta|\sigma^2) = N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$

$\Rightarrow$ posterior distribution

$$p(\theta, \sigma^2 | \vec{y}_n) = p(\theta | \sigma^2, \vec{y}_n)\, p(\sigma^2 | \vec{y}_n) \quad \text{where}$$

- $p(\theta|\sigma^2, \vec{y}_n) = N\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right)$
- $p(\sigma^2|\vec{y}_n) = IG\left(\frac{\nu_n}{2}, \frac{\nu_n}{2}\sigma_n^2\right)$

- For linear regression we have

$$Y_i \mid X_i, \beta, \sigma^2 \overset{\text{iid}}{\sim} N_1(\beta^T X_i, \sigma^2) \quad \forall i = 1, \ldots, n$$

or, in matrix notation,

$$Y_1, \ldots, Y_n \mid X_1, \ldots, X_n, \beta, \sigma^2 \sim \underbrace{N_n(\mathbf{X}\beta, \sigma^2 I_n)}_{\text{Likelihood}} \quad (*)$$

- As in the univariate normal case, we'll obtain different models depending on our choice of priors for β and σ^2

Bayesian linear regression model #1 (semi-conjugate)

- Likelihood is $(*)$
- $p(\beta) \sim N_d(\beta_0, \Sigma_0)$ ⟵ Prior for β
- $p(\sigma^2) \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right)$ ⟵ Prior for σ^2

$\Longrightarrow$ Leads to full conditionals

- $p(\beta \mid \vec{y}_n, \mathbf{X}, \sigma^2) = N_d(\beta_n, \Sigma_n)$
- $p(\sigma^2 \mid \vec{y}_n, \mathbf{X}, \beta) = IG\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_0^2 + SSR(\beta)}{2}\right)$

where

- $\Sigma_n = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)^{-1}$
- $\beta_n = \Sigma_n\left(\Sigma_0^{-1}\beta_0 + \frac{1}{\sigma^2}\mathbf{X}^T\vec{y}_n\right)$

- **Derivation of $p(\beta \mid \vec{y}_n, \mathbf{X}, \sigma^2)$.**

For notational simplicity, let $y = \vec{y}_n$ and $X = \mathbf{X}$.

$$p(\beta \mid y, X, \sigma^2) \propto p(y, X, \sigma^2 \mid \beta)\, p(\beta)$$

$$= p(y \mid X, \sigma^2, \beta)\, p(X, \sigma^2 \mid \beta)\, p(\beta)$$

Goes away b/c X is fixed and σ^2 is indep. of β

$$= p(y \mid X, \sigma^2, \beta)\, p(\beta)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right] (2\pi)^n \det(\Sigma_0)^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\beta - \beta_0)^T \Sigma_0^{-1}(\beta - \beta_0)\right]$$

$$\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}(y - X\beta)^T(y - X\beta) + (\beta - \beta_0)^T \Sigma_0^{-1}(\beta - \beta_0)\right)\right]$$

$$\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}\beta^T X^T X \beta - \frac{2}{\sigma^2}\beta^T X^T y + \beta^T \Sigma_0^{-1}\beta - 2\beta^T \Sigma_0^{-1}\beta_0\right)\right]$$

$$= \exp\left[\beta^T\left(\frac{1}{\sigma^2}X^T y + \Sigma_0^{-1}\beta_0\right) - \frac{1}{2}\beta^T\left(\frac{1}{\sigma^2}X^T X + \Sigma_0^{-1}\right)\beta\right]$$

By completing the square, one finds that the latter expression is proportional to a $N_d(\beta_n, \Sigma_n)$ density.

- Derivation of $p(\sigma^2 \mid \vec{Y}_n, \mathbf{X}, \beta)$ proceeds similarly (Hoff, p.155)
- Full conditionals $\Longrightarrow$ Gibbs sampler:
  - Initialize $\beta^{(0)}, \sigma^{2(0)}$.
  - for $1 \leq i \leq S$:
    - Compute $C = \left(\Sigma_0^{-1} + \frac{1}{\sigma^{2(i-1)}} X^T X\right)^{-1}$ and $m = C\left(\Sigma_0^{-1}\beta_0 + \frac{1}{\sigma^{2(i-1)}} X^T y\right)$
    - Sample $\beta^{(i)} \sim N_d(m, C)$
    - Compute $SSR(\beta^{(i)})$
    - Sample $\sigma^{2(i)} \sim IG\left(\frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + SSR(\beta^{(i)})}{2}\right)$
  - Return $\left\{\left(\beta^{(i)}, \sigma^{2(i)}\right) : i = 0, \ldots, S\right\}$.

Bayesian linear regression model #2 (g-priors)

- Choosing prior parameters $\beta_0, \Sigma_0, \nu_0, \sigma_0^2$ in model #1 can be challenging. For example the prior covariance $\Sigma_0 \in \mathbb{R}^{d \times d}$ for β has $\frac{d(d+1)}{2}$ parameters that need to be specified.
- Similarly to the univariate normal case, we can use a different model in which the prior for β depends on σ^2. Specifically:

- Zellner's g-prior is
  - $p(\sigma^2) = IG\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right)$ (same as before)
  - $p(\beta|\sigma^2) = N_d(0, g\sigma^2(X^TX)^{-1})$

  for some $g>0$. In this case, the posterior is

$$p(\beta, \sigma^2 | \vec{y}_n, \mathbf{X}) = p(\beta|\sigma^2, \vec{y}_n, \mathbf{X})\, p(\sigma^2|\vec{y}_n, \mathbf{X})$$

  where

- $p(\beta|\sigma^2, \vec{y}_n, \mathbf{X}) = N_d\left(\frac{g}{g+1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y, \frac{g\sigma^2}{g+1}(\mathbf{X}^T\mathbf{X})^{-1}\right)$

- $p(\sigma^2 | \vec{y}_n, \mathbf{X}) = IG\left(\frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + SSR_g}{2}\right)$

where $SSR_g = \vec{y}_n^T \left(I_n - \frac{g}{g+1} \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\vec{y}_n$

- In this model, we can sample independently $\Rightarrow$ no MCMC
- A common choice for g is the sample size, $g = n$.

Posterior inference

- So far, we have:

(i) Collected data $\{(x_i, y_i) : i = 1, \dots, n\}$

(ii) Assumed* $y_i = \beta^T x_i + \varepsilon_i$ w/ $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

(iii) Set up a Bayesian model and obtained samples $\{(\beta^{(i)}, \sigma^{2(i)})\}$

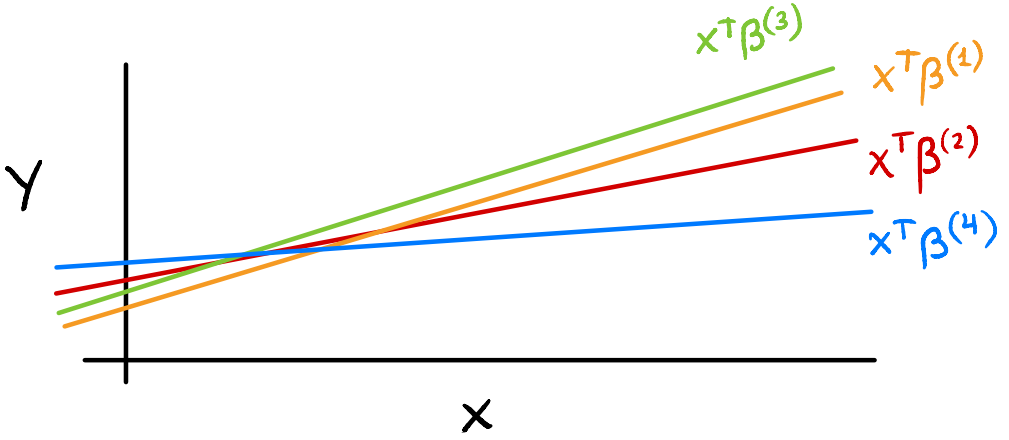
What to do with samples $\{(\beta^{(i)}, \sigma^{2(i)}) : i = 1, \dots, S\}$?

*Side note: Often, one adds a column of ones to $\mathbf{X}$ and a coefficient β_0 to $(\beta_1, \dots, \beta_d)$ to capture an intercept:

$$y = \beta^T x + \varepsilon = (\beta_0, \beta_1, \dots, \beta_d) \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} + \varepsilon = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d + \varepsilon$$

Estimating the mean function

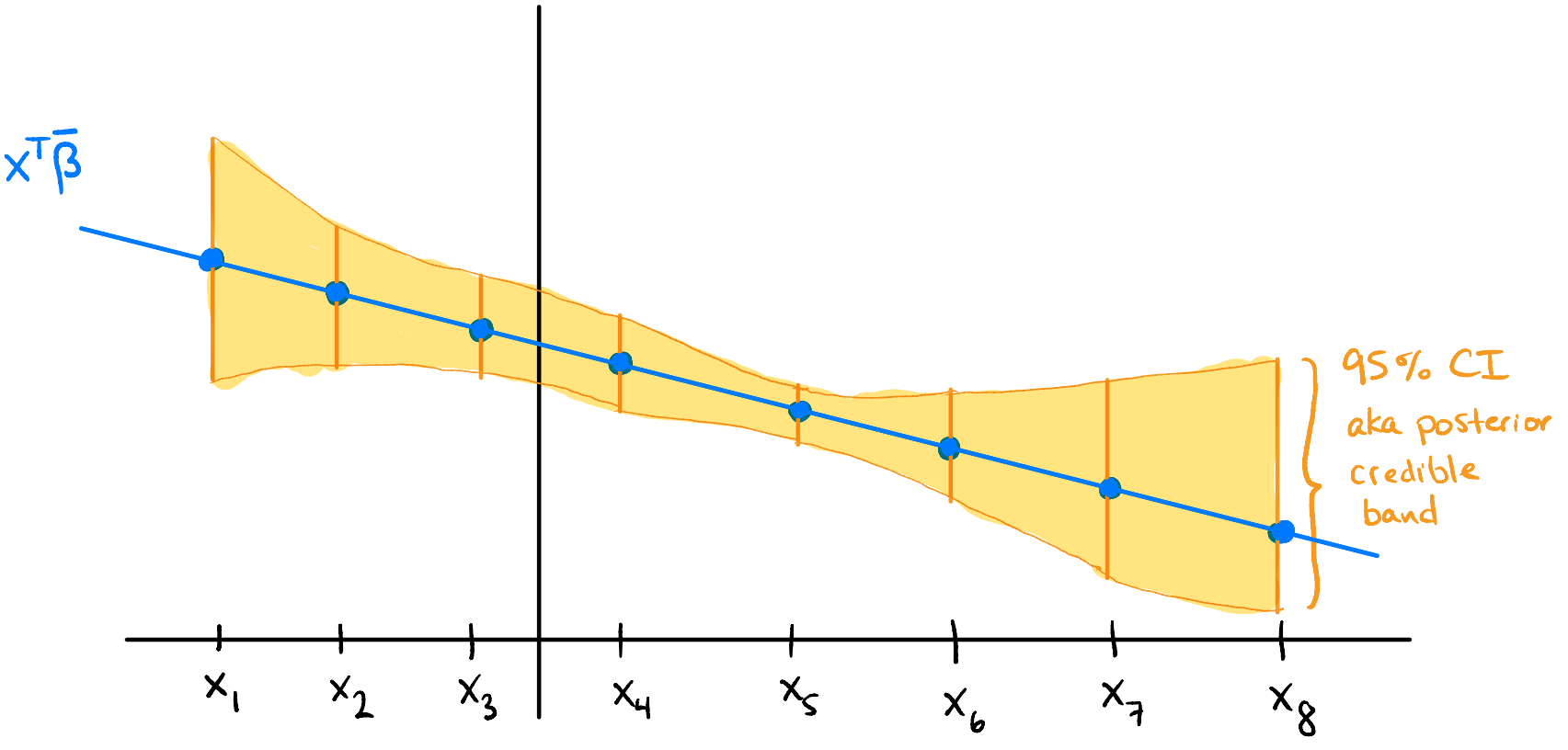
- For each $\beta = (\beta_0, \beta_1, \ldots, \beta_d) \in \mathbb{R}^{d+1}$, the function $m_\beta(x) = \beta^T x$ defines a plane in $\mathbb{R}^{d+2}$
- $m_\beta(x)$ is the mean of $y = \beta^T x + \varepsilon$ when $\mathbb{E}[\varepsilon] = 0$.
- Each sample $\beta^{(i)}$ from our posterior gives a mean estimate:



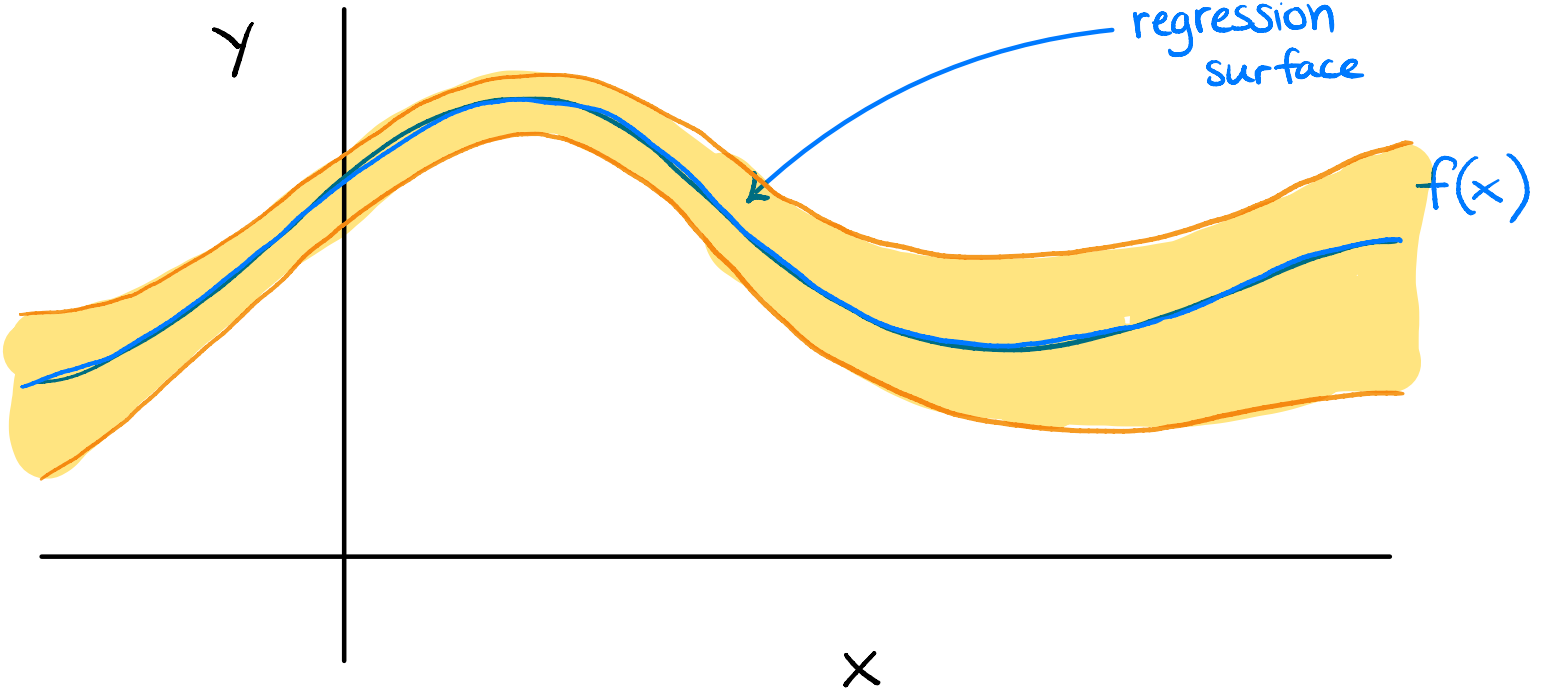
- Using our samples $\beta^{(1)}, \ldots, \beta^{(S)}$, we can compute, e.g.,
  - The sample mean:

$$\bar{\beta} = \frac{1}{S} \sum_{i=1}^{S} \beta^{(i)}$$

  - 95% confidence intervals for each x in some grid:
    - for x in grid of x-values:
      - Compute $x^T \beta^{(i)} \; \forall i$
      - Take 95% confidence interval for $\{ x^T \beta^{(i)} : i = 1, \ldots, S \}$


$x^T\bar{\beta}$
95% CI
aka posterior
credible
band
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8

- More generally, when $y = f(x) + \varepsilon$ w/ $\mathbb{E}[\varepsilon] = 0$, want to estimate the <u>mean function</u> $\mathbb{E}[y|x] = f(x)$.



Posterior prediction

- Let $\mathbf{X}$ and $\vec{y}_n$ be the previously observed data.

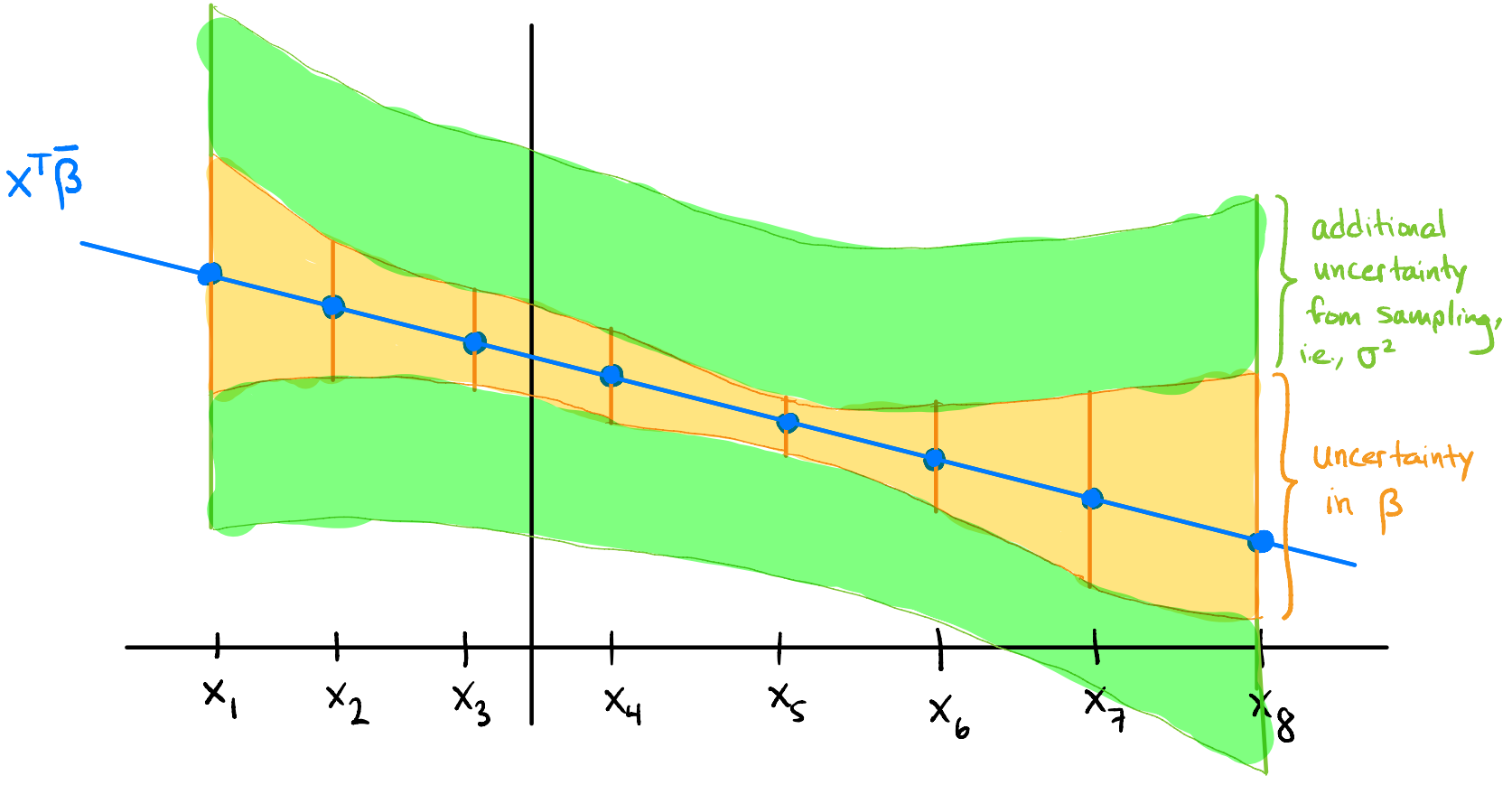
  For new y_{n+1} and x_{n+1},

$$p(y_{n+1} \mid x_{n+1}, \mathbf{X}, \vec{y}_n) = \int p(y_{n+1} \mid \beta, \sigma^2, x_{n+1}, \mathbf{X}, \vec{y}_n)\, p(\beta, \sigma^2 \mid x_{n+1}, \mathbf{X}, \vec{y}_n)\, d\beta\, d\sigma^2$$

$$= \int p(y_{n+1} \mid \beta, \sigma^2, x_{n+1})\, p(\beta, \sigma^2 \mid \mathbf{X}, \vec{y}_n)\, d\beta\, d\sigma^2$$

where $p(y_{n+1} \mid \beta, \sigma^2, x_{n+1}) = N(\beta^T x_{n+1}, \sigma^2)$

- Assuming we've already collected samples $\{(\beta^{(i)}, \sigma^{2(i)})\}_{i=1}^{S}$, e.g., via Gibbs, we can sample from PPD via
  - for $i = 1, \ldots, S$:
    - Draw $y^{(i)} \sim N(x_{n+1}^T \beta^{(i)}, \sigma^{2(i)})$
  - Return PPD samples $y^{(i)} \sim p(y_{n+1} \mid x_{n+1}, \beta, \sigma^2)$
- Similar to the mean function, we can do the above over a grid of x values:
  - for $x \in \{x_1, \ldots, x_K\}$:
    - for $i = 1, \ldots, S$
      - $y^{(i)} \sim N(x^T \beta^{(i)}, \sigma^{2(i)})$


$x^T\bar{\beta}$
additional uncertainty from sampling, i.e., σ^2
uncertainty in β
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8

Model selection

- In many applications, the number of covariates $x_1, \ldots, x_d$ is (very) large, e.g., genomics ($d \approx 20{,}000$). However, often only a (small) subset of the covariates are actually related to the response y.
- Including only the "true" covariates in our model of y often yields simpler, more interpretable models that provide better predictions and statistical properties.
- In a linear model we might have $y = \beta^T x + \varepsilon$ with some of the coefficients $\beta_j = 0$, so x_j unrelated to y.

Backward selection (OLS-based)

- One non-Bayesian approach based on t-statistics

$$t_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 (\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \quad \text{where} \quad \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{y}_n$$

is to take the following backwards elimination approach:

- Start with full $n \times d$ data matrix $\mathbf{X}$ and threshold t_*
- Set $m = \{1, \ldots, d\}$
- While $\exists j$ such that $|t_j| < t_*$:
  - find $j_{min} = \min_j |t_j|$ and remove j_{min} from m
  - compute t-statistics for new $\mathbf{X}$
- Return m

- This procedure has several problems; for example, it can identify many false positives (see Hoff, Section 9.3).

The general case

- We have a finite set of possible models M
- Goal: Given data y, compute $p(m|y) \quad \forall m \in M$.
- Assume each model $m \in M$ is parametrized by θ_m.

$$p(m|y) = \frac{p(y|m)p(m)}{p(y)} \leftarrow \text{prior on models}^*$$

$$= \frac{p(m)\int p(y|\theta_m)p(\theta_m)\,d\theta_m}{p(y)}$$

* Since θ_m defines the model m, $p(y|\theta_m, m) = p(y|\theta_m)$ and the prior on models m is equivalent to putting a prior on θ_m, i.e., $p(\theta_m) = p(m)$.

- If we want to compare two models m_1 and m_2, we can look at their <u>posterior odds</u>

$$\frac{p(m_1|y)}{p(m_2|y)} = \underbrace{\frac{p(y|m_1)}{p(y|m_2)}}_{\text{Bayes factor}} \times \underbrace{\frac{p(m_1)}{p(m_2)}}_{\text{Prior odds}}$$

- In linear regression, $y = \beta^T x + \varepsilon$, a models m correspond to subsets of $\{1, \ldots, d\}$, i.e.,

$$m = \{ j : \beta_j \neq 0 \}$$

One (of many) Bayesian approaches

- To allow for the possibility that $\beta_j = 0$ for some j, let $\beta_j = z_j b_j$ where $z_j \in \{0, 1\}$ so that

diagonal matrix $\begin{pmatrix} z_1 & & 0 \\ & z_2 & \\ 0 & & \ddots \\ & & & z_d \end{pmatrix}$

$$y = z_1 b_1 x_1 + \cdots + z_d b_d x_d + \varepsilon = \beta^T Z x + \varepsilon$$

So $z_j = 0$ means x_j is not in the model.

$\Rightarrow$ Model selection amounts to identifying which $z_j = 0$.

$\Rightarrow$ There are 2^d possible models, e.g., when $d = 3$:

$$(z_1, z_2, z_3) \in \{(0,0,0), (0,0,1), (0,1,0), (1,0,0), (0,1,1), (1,0,1), (1,1,0), (1,1,1)\}$$

- The posterior over models is

$$p(z \mid \mathbf{X}, y) = \frac{p(\vec{y}_n, \mathbf{X} \mid z)\, p(z)}{p(\vec{y}_n, \mathbf{X})} \overset{\text{since } \mathbf{X} \text{ treated as fixed}}{=} \frac{p(\vec{y}_n \mid \mathbf{X}, z)\, p(z)}{\sum_{\tilde{z}} p(\vec{y}_n \mid \mathbf{X}, \tilde{z}) p(\tilde{z})}$$

where $p(z)$ is our prior over models, e.g., Uniform on the finite set $\{(z_1, \ldots, z_d) : z_j \in \{0,1\}\} = \{0,1\}^d$.

- Usually take $p(z) = \frac{1}{2^d}$ (all models have the same prior probability).

$$\Longrightarrow p(z \mid \mathbf{X}, y) = \frac{p(y \mid \mathbf{X}, z)}{\sum_{\tilde{z}} p(y \mid \mathbf{X}, \tilde{z})}$$

- With Zellner's g-prior we can get a closed-form expression for $p(y \mid \mathbf{X}, z)$, which allows us to compute the posterior probability $p(z \mid y, \mathbf{X})$ for every model z (see Hoff, page 165).

Bayesian model averaging

- Rather than select a single model, we can average predictions over all models.
- If a model is more plausible, then its prediction should count more.
- Formally, letting $y = \vec{y}_n$ and $X = \mathbf{X}$,

$$p(y_{n+1} \mid x_{n+1}, y, X) = \sum_{m \in M} p(y_{n+1} \mid m, x_{n+1}, y, X)\, p(m \mid x_{n+1}, y, X)$$

Since x_{n+1} treated as fixed

$$= \sum_{m \in M} p(y_{n+1} \mid m, x_{n+1}, y, X)\, p(m \mid y, X)$$

$$= \sum_{m \in M} p(m \mid y, X) \int p(y_{n+1} \mid \theta_m, m, x_{n+1}, y, X) p(\theta_m \mid m, X, y)\, d\theta_m$$

↕ Because θ_m completely determines the model

$$= \sum_{m \in M} p(m \mid y, X) \int p(y_{n+1} \mid \theta_m, x_{n+1})\, p(\theta_m \mid m, X, y)\, d\theta_m$$

- In linear regression, if we have samples $\{\beta^{(i)}\}_{i=1}^{S}$:

$$\hat{\beta}_{BMA} = \frac{1}{S} \sum_{i=1}^{S} \beta^{(i)}.$$

Posterior inclusion probabilities

- Another useful quantity to look at is the posterior inclusion probability:

$$p(z_j = 1 \mid \vec{y}_n, \mathbf{X}) = \sum_{\{z : z_j = 1\}} p(z \mid \vec{y}_n, \mathbf{X}).$$

End Ch. 9