Определение достаточного объема выборки см. в Постановке задачи:

Рассмотрим правдоподобие выборки $\mathfrak{D}_{\mathcal{L}_m}$:

$$L(\mathfrak{D}_{\mathcal{T}_m}, \mathfrak{D}_{\mathcal{L}_m}) = \prod_{y,\mathbf{x}\in\mathfrak{D}_{\mathcal{T}_m}} f(y, \mathbf{x}, \mathbf{w}).$$

Рассмотрим логарифм правдоподобия выборки $\mathfrak{D}_{\mathcal{L}_m}$:

$$l(\mathfrak{D}_{\mathcal{T}_m}, \mathbf{w}) = \sum_{y,\mathbf{x}\in\mathfrak{D}_{\mathcal{T}_m}} \log f(y, \mathbf{x}, \mathbf{w}).$$

Будем рассматривать ожидаемое значение функции l:

$$\bar{l}(\mathfrak{D}) = \mathop{\mathsf{E}}_{y,\mathbf{x}\in\mathfrak{D}} l(\{y, \mathbf{x}\}, \mathbf{w}).$$

Рассмотрим ожидаемое значение логарифма правдоподобия по разным обучающим выборкам $\mathfrak{D}_{\mathcal{L}_m}$ размера m^*:

$$l(m^*) = \mathop{\mathsf{E}}_{\mathfrak{D}_{\mathcal{L}_m}} \bar{l}(\mathfrak{D}_{\mathcal{L}_m}).$$

Будем считать, что объем выборки достаточный, если:

$$\forall m_1, m_2 > m^* \quad |l(m_1) - l(m_2)| < \varepsilon,$$

где ε - достаточно малое пороговое значение.

Я использовал выборку Diabetes для задачи линейной регрессии из UCI репозитория, получилась следующий график функции l(m):

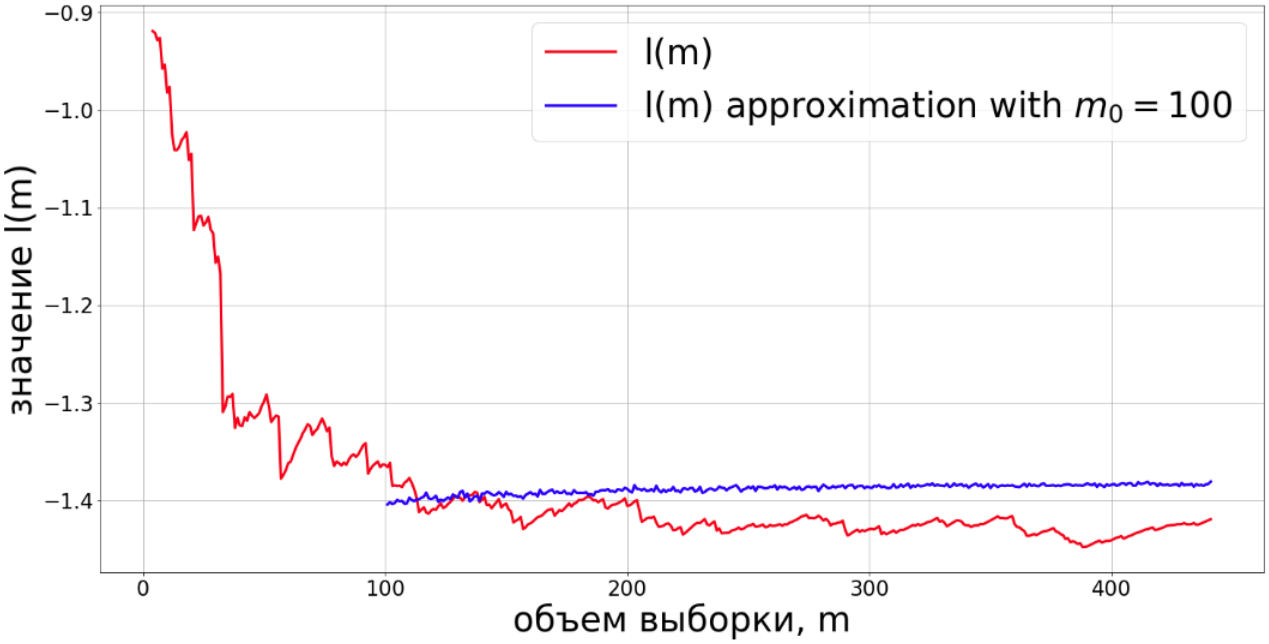
Я попытался построить аппроксимацию функции l(m), используя только первые m_0 элементов выборки следующим образом:

1) Используем аппроксимацию распределения вектора параметров для $m > m_0$:

$$\hat{\mathbf{w}} \sim \mathcal{N}(\mathbf{m}, \frac{m}{m^*} \mathbf{I}^{-1}(\mathfrak{D}_{m^*})).$$

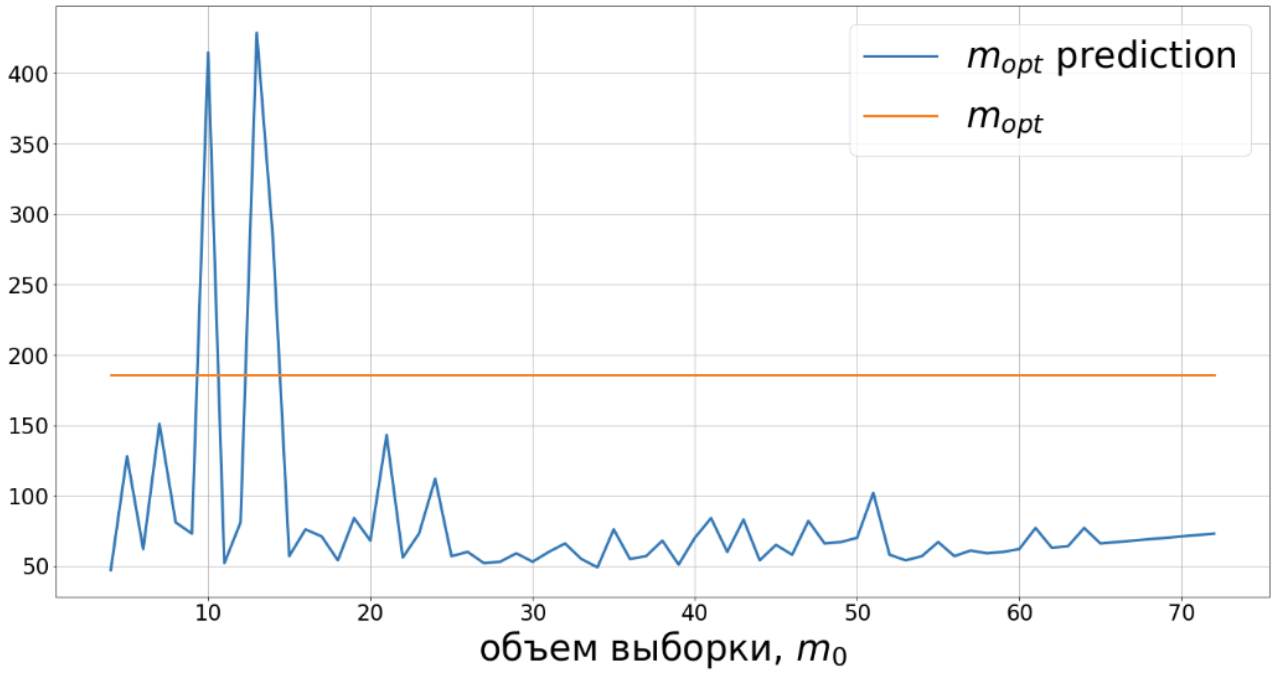
2) Много раз сэмплируем w из заданного распределения и считаем $l_w(m)$ по первым m_0 элементов, затем усредняем по всем w: $l(m) = MEAN(l_w(m))$.

Например, для m_0=100 получается следующая аппроксимация:



Тут видна проблема: функция l(m) очень зависит от выборки, на основе которой она посчитана, и, опираясь только на первые m_0 элементов, не получается построить нормальную аппроксимацию l(m).

Если считать m^* по аппроксимации l(m), получается следующий график:



Судя по всему, небходим либо другой способ аппроксимации функции l(m), либо другой критерий достаточности объема выборки.