CONTRIBUTEURS

ACT-2000 Analyse statistique des risques actuarielles

aut., cre. Alec James van Rassel

Analyse statistique des risques actuariels

Vraisemblance

On peut voir la fonction de densité $f(x;\theta)$ comme étant une fonction du paramètre inconnu θ avec x fixé ; ceci est la fonction de vraisemblance $\mathcal{L}(\theta;x)$.

Qualité de l'estimateur

La première section traite de «**estimateurs ponctuels** ». C'est-à-dire, on produit une seule valeur comme notre meilleur essai pour déterminer la valeur de la population inconnue. Intrinsèquement, on ne s'attend pas à ce que cette valeur (même si c'en est une bonne) soit la vraie valeur exacte.
Une hypothèse plus utile à des fins d'interprétation est plutôt un **estimateur par intervalle** ; au lieu d'une seule valeur, il retourne un intervalle de valeurs plausibles qui peuvent toutes être la vraie valeur. Le type principal d'*estimateur par intervalle* est *l'intervalle de confiance* traité dans la deuxième sous-section.

Estimation ponctuelle

Lorsque nous avons un estimateur $\hat{\theta}$ pour un paramètre inconnu θ on espère que, **en moyenne**, ses erreurs de prévision seront nulles. On peut alors trouver $\mathrm{E}[\hat{\theta}|\theta]$; soit, l'espérance de l'estimateur lorsque θ est la vraie valeur du paramètre. Par la suite, on calcule son **biais** $\mathrm{B}(\hat{\theta})$ dans la prévision de cette vraie valeur du paramètre :

Biais d'un estimateur

$$\mathrm{B}(\hat{\theta}) = \mathrm{E}[\hat{\theta}|\theta] - \theta$$

Cependant, le biais n'indique pas la variabilité de l'estimateur $\hat{\theta}$ dans sa prévision. Ceci est plutôt calculé avec la variance de l'estimateur $\mathrm{Var}(\hat{\theta})$. Nous pouvons alors définir la **borne inférieure de Cramèr-Rao** de la variance de l'estimateur $\mathrm{Var}(\hat{\theta})$ avec **la matrice d'information de Fisher** $I(\theta)$:

Borne inférieure Cramèr-Rao

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{n\mathrm{E}\left[\left(\frac{\partial \ln}{\partial\theta} f(x;\theta)\right)^2\right]}, \quad \text{où } I(\theta) = \mathrm{E}\left[\left(\frac{\partial \ln}{\partial\theta} f(x;\theta)\right)^2\right]$$

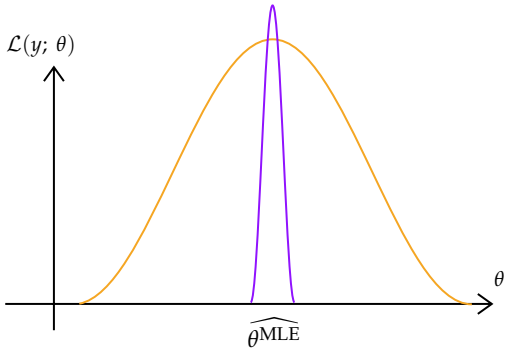
Cette borne est rarement comprise et sur la base de ce vidéo et ce vidéo je me lance dans l'explication de l'intuition derrière cette borne. Si vous ne comprenez pas, allez les regarder puisqu'elle va régulièrement réapparaître plus tard dans le bac.

Premièrement, on définit l'utilité des deux premières dérivées :

$\frac{\partial}{\partial\theta}\mathcal{L}(\theta)$: Représente le « rate of change » *(angl.)* de la fonction.

$\frac{\partial^2}{\partial\theta^2}\mathcal{L}(\theta)$: Représente la concavité de la fonction ; on peut y penser comme sa « curve ».

On sait que la dérivée sera de 0 au point $\hat{\theta}^{\mathrm{EMV}}$, mais juste passé ce point elle sera négative ! On ajoute alors un négatif pour évaluer la décroissance pareillement à l'augmentation de la dérivée. Le twist est que beaucoup de fonctions peuvent avoir le même point où elles sont maximisées, mais avoir l'air complètement différents.



Clairement, la courbe en mauve aura plus de points près de $\hat{\theta}^{\mathrm{EMV}}$ que la courbe en orange. On évalue cette différence avec la deuxième dérivée, soit la concavité ou « curve ». Ce faisant, la deuxième dérivée permet d'être plus certain d'avoir le bon estimateur. Il est alors logique que la variance ne puisse pas être moins que l'estimateur du maximum de vraisemblance évalué au point où la concavité est maximisée.

Finalement, on veut comprendre pourquoi $1/$« curve » et non juste « curve ». On déduit que plus la concavité est élevée, alors plus la variance sera faible. Il aura forcément moins de points près de $\hat{\theta}^{\mathrm{EMV}}$ et donc :

$$\mathrm{Var}(\hat{\theta}^{\mathrm{EMV}}) \overset{\text{dépend}}{\sim} \frac{1}{\text{« curve »}}$$

On observe alors que la limite lorsque la « curve » tend vers l'infini implique une variance nulle. On dit donc que la distribution de l'estimateur est "asymptotiquement normale" tel que $\hat{\theta}^{\mathrm{EMV}} \overset{a.s.}{\Rightarrow} \mathcal{N}\left(\theta, \left(\boldsymbol{I}(\theta)\right)^{-1}\right)$ où a.s. veut dire asymptotiquement.

Il s'ensuit alors que **l'efficacité d'un estimateur** est le ratio de la borne Cramèr-Rao sur la variance de l'estimateur :

Efficacité d'un estimateur

$$\text{eff}(\theta) = \frac{\text{Var}(\hat{\theta})^{\text{Rao}}}{\text{Var}(\hat{\theta})} = \frac{1}{n\text{E}\left[\left(\frac{\partial \ln}{\partial \theta} f(x;\theta)\right)^2\right] \text{Var}(\hat{\theta})}$$

Si ce ratio est de 1, $\text{eff}(\theta) = 1$, alors l'estimateur est à la borne. On dit qu'il est un estimateur « **efficient** » *(angl.)*. Du fait, parmi les estimateurs sans biais il doit être celui avec la **variance minimale** ; on dit que l'estimateur est le « **Minimum Variance Unbiased Estimator (MVUE)** » *(angl.)*.
De plus, on peut généraliser cette formulation pour obtenir l'efficacité relative d'un estimateur comparativement à un autre :

Efficacité relative

Soit les estimateurs non biaisés $\hat{\theta}_n$ et $\tilde{\theta}_n$, l'efficacité de $\hat{\theta}_n$ relativement à $\tilde{\theta}_n$ est :

$$\text{eff}(\hat{\theta}_n, \tilde{\theta}_n) = \frac{\text{Var}(\hat{\theta}_n)}{\text{Var}(\tilde{\theta}_n)}$$

Si $\text{eff}(\hat{\theta}_n, \tilde{\theta}_n) < 1$ alors $\hat{\theta}_n$ est plus efficace et vice-versa.
Nous pouvons également évaluer si un estimateur est cohérent, ou converge, avec des très grands échantillons ; un estimateur $\hat{\theta}$ est **consistent** *(angl.)* si la probabilité qu'il diffère de la vraie valeur du paramètre θ par une erreur ϵ près de 0 tend vers 0 alors que la taille de l'échantillon n tend vers l'infini :

Convergence (consistency) d'un estimateur

$$\lim_{n\to\infty} \Pr(\left|\hat{\theta} - \theta\right| > \epsilon) = 0, \quad \epsilon > 0$$

Ce critère peut être rencontré lorsque l'estimateur $\hat{\theta}$ est **asymptotiquement sans biais *et*** que la **variance de l'estimateur** tend vers 0. D'ailleurs, nous avons déjà raisonné ceci avec la borne inférieure Cramèr-Rao.
On définit proprement ce qu'est un estimateur **asymptotiquement sans biais** :

Estimateur asymptotiquement sans biais

$$\lim_{n\to\infty} \text{B}(\hat{\theta}) = 0$$

Donc si $\lim_{n\to\infty} \text{Var}(\hat{\theta}) = 0$ et que $\lim_{n\to\infty} \text{B}(\hat{\theta}) = 0$ alors l'estimateur est **consistent**. Cependant, l'inverse n'est pas vrai ; un estimateur qui est **consistent** n'implique pas que la variance et le biais tendent vers 0.

Malgré la nature plaisante de la convergence d'un estimateur, beaucoup d'estimateurs ont cette propriété. Nous voulons alors une mesure qui n'indique pas seulement qu'un estimateur arrive près de la bonne valeur souvent *(alias, une très petite variance)*, mais qu'il est mieux que d'autres estimateurs. De plus, dût à la sélection arbitraire de l'erreur ϵ pour la *consistency* d'un estimateur, il est possible de malicieusement la sélectionner pour faire parler les données comme on le souhaite.
Nous définissons alors l'**Erreur Quadratique Moyenne** (EQM), ou **Mean Squared Error (MSE)**, permettant de comparer les différents estimateurs ayant tous une bonne *consistency* en assurant une cohérence d'interprétation.

Erreur Quadratique Moyenne (Mean Squared Error)

$$\text{MSE}_{\hat{\theta}}(\theta) = \text{E}[(\hat{\theta} - \theta)^2 | \theta] \Leftrightarrow \text{Var}(\hat{\theta}) + \left[\text{B}(\hat{\theta})\right]^2$$

En combinant tous ces critères, le meilleur estimateur est alors l'estimateur **sans biais** ayant la **plus petite variance** possible parmi tous les estimateurs *sans biais* possible ; c'est-à-dire, le **Uniformly Minimum Variance Unbiased Estimator *(UMVUE)***.

Estimation par intervalles

Un type d'estimateur par intervalle est l'**intervalle de confiance** :

Intervalle de confiance

Soit le paramètre à estimer θ, alors nous sommes confiants à un niveau de $100(1-\alpha)\%$ qu'il est contenu entre (L, U).
De façon équivalente, nous sommes confiant à un seuil de $\alpha\%$ qu'il est contenu entre (L, U) :

$$\theta \in [L, U].$$

Nous pouvons alors dire que $\Pr(L \leq \theta \leq U) \geq (1-\alpha)$ pour tout θ.

Par exemple, dans le cas d'une population avec distribution normale et moyenne μ inconnue, on a la moyenne échantillonnale $\bar{x}$ (qui est l'estimateur *MVUE*).

Intervalle de confiance sur la moyenne *(distribution normale)*

Nous sommes confiants à un niveau de $100(1-\alpha)\%$ que :

$$\mu \in \left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Construction d'estimateurs

Dans la section précédente, on évalue les méthodes pour évaluer la **qualité** de l'estimateur. Cependant, comment obtenons-nous des estimateurs pour les évaluer ?
Plusieurs méthodes existent pour établir des estimateurs, de plus plusieurs méthodes existent pour estimer des paramètres. La méthode vu dans le cadre du cours de statistique est la **méthode fréquentiste**, le cours de mathématiques IARD 1 (ACT-2005) présente **l'estimation bayésienne**.
Avant de le faire, nous présentons quelques concepts :

échantillon aléatoire : Échantillon d'observations indépendantes provenant de la même distribution paramétrique (identiquement distribué) ; c'est-à-dire, un échantillon **(iid)**.

k-ème moment centré à 0 : $\mu'_k = \mathrm{E}[X^k]$.

$100g^{\text{ème}}$ **pourcentile :** $\pi_g(\theta) = F_\theta^{-1}(g)$.

Les deux premiers estimateurs ci-dessous sont les plus faciles à obtenir, mais sont aussi les moins performants puisqu'ils n'utilisent que quelques traits des données au lieu de l'entièreté des données comme la troisième méthode.
Cette distinction devient particulièrement importante dans le cas d'une distribution avec une queue lourde à la droite (Pareto, Weibull, etc.) où il devient plus essentiel de connaître les valeurs extrêmes pour bien estimer le paramètre de forme (α pour une Pareto).
Un autre désavantage est que les deux premières méthodes nécessitent que les données proviennent toutes de la même distribution, autrement les moments et quantiles ne seraient pas clairs.
Finalement, sous les deux premières méthodes la décision de quels moments et percentiles à utiliser est arbitraire.

Méthode des moments (MoM)

Soit un échantillon aléatoire de taille n (iid), on pose $\hat{\mu}'_k = \mu'_k$.

Estimation de θ par la méthode des moments

L'estimation de θ est alors toute solution des p équations :
$$\mu'_k(\theta) = \hat{\mu}'_k, \quad k = 1, 2, \ldots, p$$

La raison pour cet estimateur est que la distribution empirique aura les mêmes p premiers moments centrés à 0 que la distribution paramétrique.

Méthode du «Percentile Matching »

Soit un échantillon aléatoire de taille n (iid), on pose $\hat{\pi}_g(\theta) = \pi_g(\theta)$.

Estimation de θ par la méthode du «Percentile Matching »

L'estimation de θ est alors toute solution des p équations :
$$F(\hat{\pi}_{g_k}|\theta) = g_k, \quad k = 1, 2, \ldots, p$$

La raison pour cet estimateur est que le modèle produit aura p percentiles qui vont «matcher »les données.

Il peut arriver que les percentiles de distributions ne soient pas uniques, par exemple dans le cas de données discrètes lorsque le quantile recherché peut tomber entre 2 *marches* de la fonction empirique, ou mal-définis. Il est alors utile de définir une méthode d'interpolation des quantiles (bien qu'il n'en existe pas une d'officielle).

Soit le «**smoothed empirical estimate** »d'un percentile :

Smoothed empirical estimate

On utilise les statistiques d'ordre de l'échantillon $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ pour l'interpolation suivant :
$$\hat{\pi}_g = (1-h)x_{(j)} + hx_{(j+1)}, \quad \text{où}$$
$$j = \lfloor (n+1)g \rfloor \qquad \text{et} \qquad h = (n+1)g - j$$

Méthode du maximum de vraisemblance

Nous cherchons à maximiser la probabilité d'observer les données. Ceci est fait par la vraisemblance $\mathcal{L}(\theta; x)$ ou, puisque le logarithme ne change pas le maximum, la log-vraisemblance $\ell(\theta; x)$ où :

Maximum de vraisemblance

$$\mathcal{L}(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta) \qquad \text{et} \qquad \ell(\theta; x) = \sum_{i=1}^{n} \ln f(x_i; \theta)$$
et l'**estimateur du maximum de vraisemblance** de $\boldsymbol{\theta}$ est celui qui maximise la fonction de vraisemblance.

Statistiques d'ordre

Soit un échantillon aléatoire de taille n. Nous définissons la k^{e} **statistique d'ordre** $X_{(k)}$ comme étant la k^{e} plus petite valeur d'un échantillon.
Les crochets sont utilisés pour différencier la k^{e} statistique d'ordre $X_{(k)}$ de la k^{e} observation X_k.

Nous sommes habituellement intéressés au minimum $X_{(1)}$ et le maximum $X_{(n)}$.

Minimum

$$X_{(1)} = \min(X_1, \dots, X_n)$$
$$f_{X_{(1)}}(x) = n f_X(x) \big(S_X(x)\big)^{n-1}$$
$$S_{X_{(1)}}(x) = \prod_{i=1}^{n} \Pr(X_i > x)$$

Maximum

$$X_{(n)} = \max(X_1, \dots, X_n)$$
$$f_{X_{(n)}}(x) = n f_X(x) \big(F_X(x)\big)^{n-1}$$
$$F_{X_{(n)}}(x) = \prod_{i=1}^{n} \Pr(X_i \leq x)$$

De façon plus générale, on défini :

k^{e} statistique d'ordre

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!1!(n-k)!} \underbrace{[F_X(x)]^{k-1}}_{\text{observations} < k} \overbrace{f_X(x)}^{\text{observation} = k} \underbrace{[S_X(x)]^{n-k}}_{\text{observations} > k}$$

$$F_{X_{(k)}}(x) = \underbrace{\sum_{i=1}^{n} \binom{n}{i} [F_X(x)]^j [1 - F_X(x)]^{n-j}}_{\text{Probabilité qu'au moins } k \text{ des } n \text{ observations } X_k \text{ sont} \leq x}$$

Nous pouvons également définir quelques autres statistiques d'intérêt :

$R = X_{(n)} - X_{(1)}$: **L'étendue** (range) est la différence entre le minimum et le maximum d'un échantillon.

- L'utilité de l'étendue est limitée puisqu'elle est très sensible aux données extrêmes.
- Par exemple, supposons qu'on observe des données historiques de température pour le 1er septembre.
  En moyenne, la température est de $16°C$, mais nous avons un cas extrême de $-60°C$ en 1745.
  L'étendue sera de $86°C$ ce qui n'est très représentatif des données.
  Donc, dans ce contexte, la mesure n'est pas d'une très grande utilité.

$M = \frac{X_{(n)} + X_{(1)}}{2}$: **mi-étendue** (Midrange), est la moyenne entre le minimum et le maximum d'un échantillon.

- Pour comprendre ce que représente la mi-étendue, on la compare à la moyenne arithmétique.
- La moyenne arithmétique considère les données observées et calcule leur moyenne.
  Il s'ensuit qu'elle ne considère pas les chiffres qui ne sont pas observés.
- La mi-étendue considère **tous** les chiffres, observés ou non, entre la plus grande et la plus petite valeur d'un échantillon et en prend la moyenne.

Exemple sur les statistiques d'ordre

Soit un échantillon de données météorologiques $\{-30°, -24°, -7°, -23°, +5°\}$ (celsius).
Je suppose que ce sont des températures du 4 février observées lors des dernières années.

- La moyenne arithmétique ($-22.25°C$) m'intéresse, car je peux savoir, en moyenne, ce qu'est la température le 4 février.
- La mi-étendue ($-12.5°C$), tout comme l'étendue ($-35°C$), ne m'intéresse pas puisqu'elle ne prend pas en considération la vraisemblance des différentes températures.

Maintenant, je suppose que ces données sont des températures observées tout au long de l'hiver passé.

- La moyenne arithmétique ne m'intéresse pas puisqu'elle est beaucoup trop biaisée par les températures de cette même journée.
- Cependant, la mi-étendue et l'étendue me donnent maintenant une meilleure idée de la température de l'hiver.

L'important à retenir est que l'utilité des mesures dépend de la situation. Également, ceci est un exemple **très** simpliste et dans tous les cas on ne peut pas tirer de conclusions sur les températures de l'hiver à partir d'une seule journée.

Nous pouvons définir la **médiane** en termes de statistiques d'ordre :

$$\text{Med} = \begin{cases} X_{((n+1)/2)}, & \text{si n est impair} \\ \frac{X_{(n/2)} + X_{(n/2+1)}}{2}, & \text{si n est pair} \end{cases}$$

Finalement, on définit la distribution conjointe du minimum et du maximum $\forall x < y$:

$$f_{X_{(1)}, X_{(n)}}(x, y) = n(n-1)[F_X(y) - F_X(x)]^{n-2} f_X(x) f_X(y)$$

Modèles linéaires en actuariat

Régression linéaire simple

Modèle de régression linéaire simple

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Exemple de compréhension

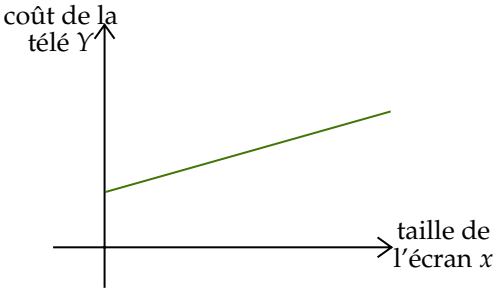
On illustre le concept et la signification des paramètres de régression avec cet exemple illustratif

Objectif On veut deviner le coût d'une télévision (télé) selon la taille de son écran.

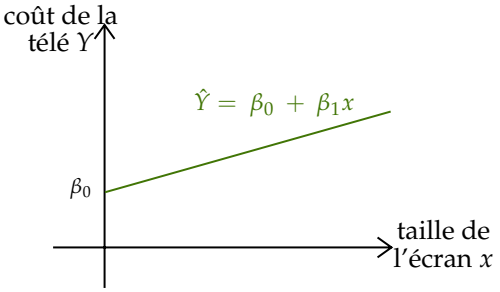
L'idée de la "régression" est de deviner, ou "prédire" du mieux qu'on peut le coût d'une télé en fonction de la taille de son écran.
Deviner le coût *exact* d'une télé *seulement* en fonction de la taille de son écran est impossible. Il y a de nombreuses raisons qui déterminent le prix d'une télé et un bon exercice est de réfléchir à ce qu'elles pourraient être. J'inclus ci-dessous une liste de quelques raisons, ou "facteurs", qui me sont survenus :

- La compagnie qui la produit (Sony vs LG, etc.).
- La résolution (4K vs 360p).
- L'année de fabrication (1990 vs 2020).
- L'endroit de l'achat (Amazon vs BestBuy, Mexique vs Canada, etc.).
- Le temps de l'année (été vs hiver, Boxing Day, etc.).

Maintenant supposons que tu joues à un jeu avec tes amis où qu'ils doivent deviner le coût d'une télé en fonction de sa taille. Ils vont probablement tous te donner des différentes réponses.
Si tu crées un modèle de prévision, il doit être systématique et toujours deviner le même prix pour la même taille d'écran—même si la prévision est erronée.
Alors, supposons que tu changes le jeu un peu et stipules que la personne qui devine le prix le plus éloigné doit prendre une gorgée de sa bière. Les réponses de tes amis vont probablement se ressembler un peu plus, mais il y a un problème qui demeure—tu veux que les prévisions soient proportionnelles à la taille de l'écran. C'est-à-dire, si ton ami devine qu'une télé de 25" coûte 100$, tu t'attends à ce qu'il devine qu'une télé de 50" coûte 200$.
La raison est qu'une régression **linéaire** *simple* est simplement une ligne droite :

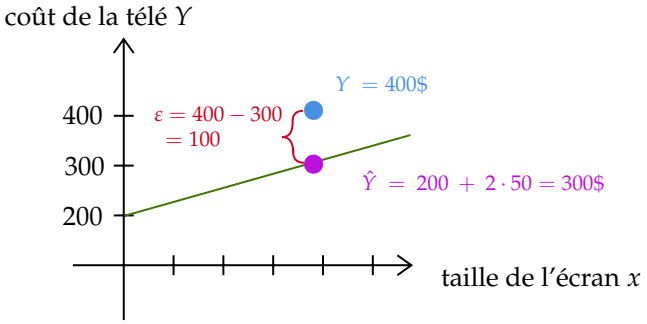


L'intuition est que ton ami se base uniquement sur la taille de l'écran comme information pour deviner le coût. Une régression **linéaire** simple applique un facteur **multiplicatif**. Il ne peut pas se dire que plus grand l'écran est grand, plus le prix va augmenter—ceci serait plutôt une régression avec un paramètre **exponentiel**.
On crée donc un facteur surnommé "paramètre". Dans le cas d'une régression linéaire simple, on a deux paramètres d'intérêts : un "niveau de base" pour le coût β_0 et un "multiplicateur" de la taille d'écran β_1 :



On suppose qu'une télé doit coûter au moins un certain prix. Ce "niveau de base" est l'intercepte sur le graphique ci-dessus surnommé l'ordonnée β_0. De ton gré, tu supposes au moins $\beta_0 = 200\$$ pour cet exemple.
Ensuite, le multiplicateur va multiplier la taille de l'écran pour obtenir un prix. Ce paramètre représente donc la pente β_1. De ton gré, tu suppose une pente de $\beta_1 = 2\$$ pour cet exemple.
Le coût (l'axe des Y) est la variable qui dépend de la taille—c'est la variable "dépendante" Y. La taille (l'axe des x) est la variable que l'on connaît indépendamment du coût—c'est la variable "indépendante" x.
Finalement la droite elle-même est le coût que le modèle devine $\hat{Y}$. Le chapeau signifie que c'est une estimation, ou "prévision".
Par exemple, le modèle devine que le prix d'une télé de 50" est de 300$; soit, $\hat{Y} = \beta_0 + \beta_1 x = 200 + (2) \cdot (50) = 300$. Selon le modèle, on estime que le coût de la télé est de 300$.

Maintenant, si tu connais le *vrai* coût réel Y tu peux mesurer à quel point tu est dans le champ. Supposons que le vrai coût est de $Y = 400\$$. Alors, l'erreur dans ta prédiction est de $\varepsilon = 400 - 300 = 100\$$.
Graphiquement :



On voit donc que $Y = \beta_0 + \beta_1 x + \varepsilon$ est un "modèle" théorique pour obtenir une variable dépendante Y en fonction de :

- Une variable indépendante x multipliée par un facteur β_1.
- Un niveau de base l'intercepte β_0.
- Une erreur aléatoire ε inconnue.