Recall: $\mu_{Y|X}$ = Regression of Y on X

$\mathbb{E}(Y|X)$ = conditional expectation.

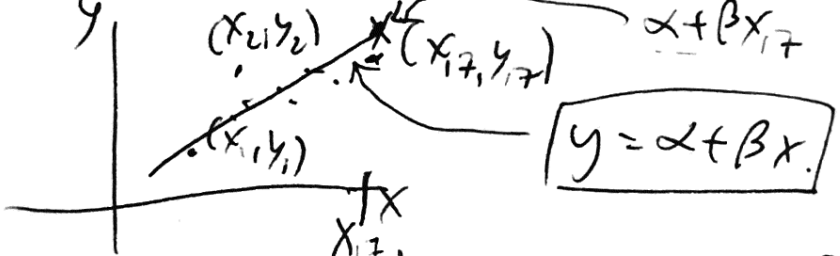
If linear, $\mu_{Y|X} = \alpha + \beta X$; then.

$$\mu_{Y|X} = \nu + \frac{c}{\sigma^2}(x-\mu), \text{ where}$$

$\nu = \mathbb{E}(Y)$, $\mu = \mathbb{E}(x)$, $\sigma^2 = \mathrm{Var}(x)$, $c = \mathrm{Cov}(X,Y)$.

i.e. $\mu_{Y|X} = \underbrace{\nu - \mu\frac{c}{\sigma^2}}_{\alpha} + \underbrace{\frac{c}{\sigma^2}}_{\beta} x$.

Assume regression of Y on X is linear. Observe $(x_i, y_i)_{i=1}^n$



What should we optimize? What are the errors?

Each error looks like: $(\alpha+\beta x_i) - y_i$

Legendre: Optimize $\sum_{i=1}^{n} (y_i - (\alpha+\beta x_i))^2 = Q$.

To find minimum of Q, solve $\frac{\partial Q}{\partial \alpha} = 0 = \frac{\partial Q}{\partial \beta}$.

Will give "least square error" estimators $\hat{\alpha}, \hat{\beta}$.

$$\frac{\partial Q}{\partial \alpha} = \sum_{i=1}^{n} 2(y_i - (\alpha + \beta x_i))^1 (-1) = 0.$$

$$\Leftrightarrow \quad n\bar{y} - n\alpha - \beta n\bar{x} = 0$$

$$\Leftrightarrow \quad \alpha + \beta\bar{x} = \bar{y} \quad \Leftrightarrow \quad \alpha = \bar{y} - \beta\bar{x}$$

$$0 = \frac{\partial Q}{\partial \beta} = \sum_{i=1}^{n} 2(y_i - (\alpha + \beta x_i))^1 (-x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i = \alpha n\bar{x} + \beta \sum_{i=1}^{n} x_i^2$$

Let $S_{xy} := \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$.

$$S_{xx} := \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

$$S_{xy} + n\bar{x}\bar{y} = \alpha n\bar{x} + \beta(S_{xx} + n\bar{x}^2).$$

$$- n\bar{x}\bar{y} = -\alpha n\bar{x} - n\beta\bar{x}^2$$

$$S_{xy} = \beta S_{xx}, \quad \boxed{\hat{\beta} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\alpha} = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}.}$$

How good are $\hat{\alpha}$ & $\hat{\beta}$ as estimators for α & β?

Extra assumption: normal (linear) regression analysis:

$f_{Y}(y|X=x)$ = conditional density

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-(\alpha+\beta x))^2}{2\sigma^2}}$$

$\mu_{Y|X} = \alpha + \beta x$.

"

$E(Y|X)$.

With this extra assumption, can compute max likelihood estimators for $\hat{\alpha}, \hat{\beta}, \hat{\sigma}$.

Observed $(x_i, y_i)_{i=1}^{n}$.

$$L(y_1, \dots, y_n \mid x_1, \dots, x_n, \alpha, \beta, \sigma) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(y_i-(\alpha+\beta x_i))^2}{2\sigma^2}}$$

$\frac{\partial L}{\partial \alpha} = 0 = \frac{\partial L}{\partial \beta} = \frac{\partial L}{\partial \sigma}$ ⟷ same with $\log L$.

$$\log L = -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2$$

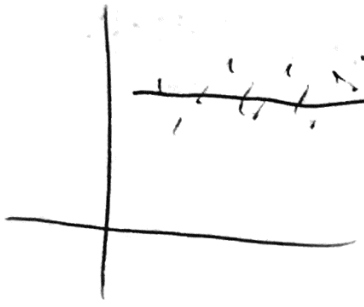
$\frac{\partial \log L}{\partial \alpha} = 0 = \frac{\partial \log L}{\partial \beta} \Rightarrow \hat{\alpha}$ & $\hat{\beta}$ as before.

$$0 = \frac{\partial \log L}{\partial \sigma} = \frac{-n\sigma^2}{\sigma^3} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2 = 0.$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - (\hat{\alpha}+\hat{\beta}x_i)\right)^2}$$

(exercise) $$= \sqrt{\frac{1}{n}\left(S_{yy} - \hat{\beta} S_{xy}\right)}$$

Let's test (under normality assumption) hypotheses on parameters, I.e. $\hat{\beta} = 0$?



$\hat{\beta}$ ← correlation

$$\frac{S_{xy}}{S_{xx}}$$

Null hypothesis (if we want to show that the variables are correlated).

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{S_{xx}}$$

y_i = observed values

Y_i = RV

Let $$\hat{B} = \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{S_{xx}}\right)(Y_i-\bar{Y}) = RV.$$

$$= \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{S_{xx}}\right) Y_i$$ ← normal

↑ Normal!

Need: mean & variance $\hat{B}$.

$$\mathbb{E}(\hat{B}) = \mathbb{E}(\hat{B} \mid X_i) = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right) \underbrace{\mathbb{E}(Y_i \mid x_i)}_{(\alpha + \beta x_i)}$$

$$= \beta \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right) x_i = \frac{\beta}{S_{xx}} \underbrace{\left[\sum_{i=1}^{n} x_i^2 - \bar{x}\, n \bar{x}\right]}_{S_{xx}}$$

$$= \beta.$$

Next: $\mathrm{Var}(\hat{B}) = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2 \underbrace{\mathrm{Var}(Y_i)}_{\sigma^2}$

$$= \frac{\sigma^2}{S_{xx}^2} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})^2}_{S_{xx}} = \boxed{\frac{\sigma^2}{S_{xx}}}.$$

Recap: $\hat{B} = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right) Y_i$ is a normal random variable with mean $\mathbb{E}(\hat{B}) = B$ & $\mathrm{Var}(\hat{B}) = \frac{\sigma^2}{S_{xx}}$.

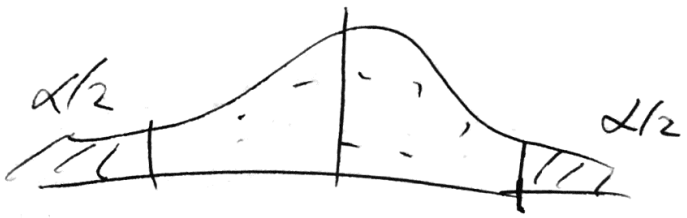
Fact: $\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2}(S_{YY} - \hat{B} S_{XY}) \sim \chi^2_{n-2}$ deg.

& $\frac{n\hat{\sigma}^2}{\sigma^2}$ is indep of $\hat{B}$.

$$\frac{Z}{\sqrt{Y/n}} = T_{n-2} = \frac{\hat{B}-\beta}{\sqrt{\sigma^2/S_{xx}}} \cdot \sqrt{\frac{\sigma^2 \quad (n-2)}{n\hat{\sigma}^2}}$$

Thm: $(\hat{B}-\beta)\cdot\sqrt{\frac{S_{xx}(n-2)}{\hat{\sigma}^2 \; n}}$ is a T-distributed RV with $n-2$ degrees of freedom.

Q: Construct a $(1-\alpha)100\%$ confidence interval for β:



$$P\left(-t_{\alpha/2,n-2} < (\hat{B}-\beta)\sqrt{\frac{(n-2)S_{xx}}{n\hat{\sigma}^2}} < t_{\alpha/2,\,n-2}\right) = 1-\alpha.$$

$$\Downarrow$$

$$\hat{B} - \_\_\_\_ < \beta < \hat{B} + t_{\alpha/2,n-2}\cdot\sqrt{\frac{n\hat{\sigma}^2}{(n-2)S_{xx}}}$$