universität wien

Statistics for Data Science, WS2023

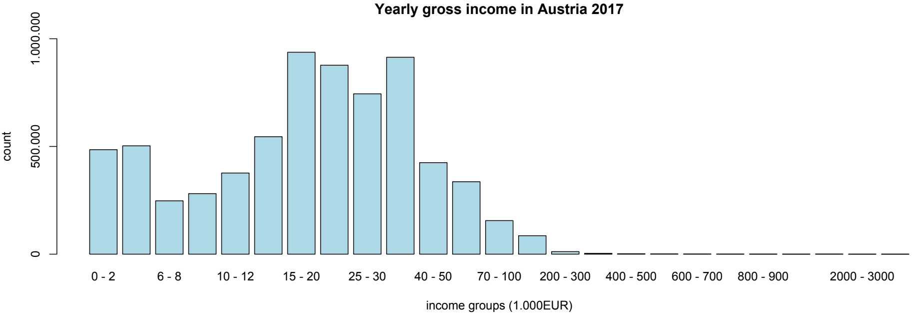
Chapter 4:

Linear Models and Model Selection

The Gaussian linear model

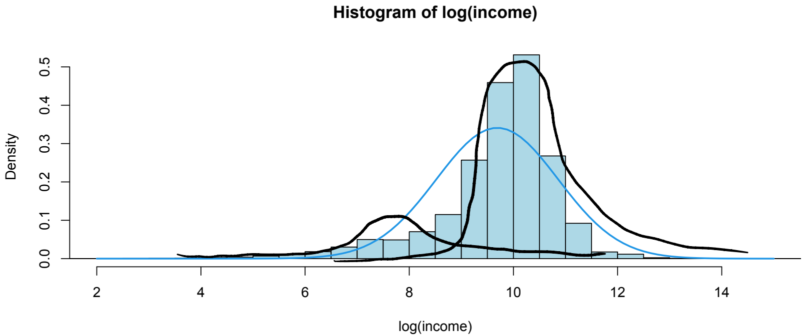
The Gaussian linear model: Motivation

Recall our income example:



What are the most important factors that influence a persons income? How large is the gender pay gap?

Often, data are (nearly) Gaussian after an appropriate transformation.



They shouldn't really be! We are looking at many different sub-populations.

The Gaussian linear model: Motivation

We want to 'explain' the (log) income using other variables, e.g., gender, age, education, etc.

$$\underset{\text{log income}}{Y = \begin{pmatrix} 8.23 \\ 11.54 \\ 10.02 \\ \vdots \\ 7.78 \end{pmatrix}} \qquad \underset{\text{intercept}}{X_{\cdot 1} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}} \qquad \underset{\text{gender}}{X_{\cdot 2} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ \vdots \\ 0 \end{pmatrix}} \qquad \underset{\text{age}}{X_{\cdot 3} = \begin{pmatrix} 37 \\ 25 \\ 62 \\ \vdots \\ 18 \end{pmatrix}}$$

Y is called the **response** or **dependent** variable.
The $X_{\cdot 1}, \ldots, X_{\cdot p}$ are called: **covariates**, **predictor-**, **regressor-**, **explanatory-**, **feature-** or **independent** variables.

universität wien

- $Y \sim N(X\beta, \sigma^2 I_n)$, $\beta \in \mathbb{R}^p$, $\sigma^2 \in (0, \infty)$. $\Theta = \mathbb{R}^p \times (0,\infty)$
- Low-dimensional case: $p < n$
- X is an $n \times p$ (non-random) design matrix with $\operatorname{rank}(X) = p$ (e.g., analysis conditional on X)

matrix with independent var.

"The mean of Y is assumed to be a linear function of our explanatory variables $X_{\cdot 1}, \ldots, X_{\cdot p}$, i.e.,

$$\mathbb{E}[Y] = X\beta = \beta_1 X_{\cdot 1} + \cdots + \beta_p X_{\cdot p} \in \mathbb{R}^n$$

or

$$\mathbb{E}[Y_i] = X_{i\cdot}\beta = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} \in \mathbb{R}.$$

"β_k is the expected change of the response variable when the regressor $X_{\cdot k}$ increases by one unit and all the others stay the same."

- $Y \sim N(X\beta, \sigma^2 I_n)$, $\beta \in \mathbb{R}^p$, $\sigma^2 \in (0, \infty)$.
- Low-dimensional case: $p < n$
- X is an $n \times p$ (non-random) design matrix with $\mathrm{rank}(X) = p$ (e.g., analysis conditional on X)

Ordinary least squares estimators:

- $\hat{\beta} := \mathrm{argmin}_{b \in \mathbb{R}^p} \|Y - Xb\|_2^2 = (X'X)^{-1}X'Y$
- $\hat{\sigma}^2 := \frac{1}{n-p}\|Y - X\hat{\beta}\|_2^2$
- $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$
- $\frac{\hat{\sigma}^2(n-p)}{\sigma^2} \sim \chi^2_{n-p}$ independent of $\hat{\beta}_n$

$$\hat{\beta} := \operatorname{argmin}_{b \in \mathbb{R}^p} \underbrace{\|Y - Xb\|_2^2}_{=:L(b)}$$

$$\nabla L(b) = \nabla \sum_{i=1}^{n} (Y_i - X_{i\cdot} b)^2 = -2 \sum_{i=1}^{n} X_{i\cdot}' (Y_i - X_{i\cdot} b) = -2X'(Y - Xb)$$

$\nabla^2 L(b) = 2X'X$ is positive definite

Normal equations: $-2X'(Y - Xb) = 0 \iff X'Xb = X'Y$

$$\Rightarrow \hat{\beta} = (\underbrace{X'X}_{p \times p})^{-1} X'Y$$

$$Y \sim N(X\beta, \sigma^2 I_n), \quad \mathbb{E}[Y] = X\beta = \sum_{k=1}^{p} \beta_k X_{\cdot k},$$

$$\hat{\beta} = (X'X)^{-1}X'Y \sim N(\beta, \sigma^2(X'X)^{-1})$$

$$e_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \swarrow k$$

Want to do statistical inference on individual effects.
E.g.: $H_0 : \beta_2 = 0$ $vs.$ $H_1 : \beta_2 \neq 0$

$$\hat{\beta}_k = e_k'\hat{\beta} \sim N(e_k'\beta, e_k'[\sigma^2(X'X)^{-1}]e_k) = N(\beta_k, \sigma^2[(X'X)^{-1}]_k)$$

$$se(\hat{\beta}_k) = \sigma\sqrt{[(X'X)^{-1}]_k}$$

where $[(X'X)^{-1}]_k$ is the k-th diagonal entry of $(X'X)^{-1}$.

One can show that

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}\sqrt{[(X'X)^{-1}]_k}} \sim t_{n-p}, \qquad \text{Student-t distribution}$$

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}\sqrt{[(X'X)^{-1}]_k}} \sim t_{n-p}, \qquad \text{Student-t distribution}$$

$$T_k := \frac{\hat{\beta}_k - b}{\hat{\sigma}\sqrt{[(X'X)^{-1}]_k}} \sim t_{n-p}$$

e.g. $b = 0$

under the null hypothesis $H_0 : \beta_k = b$. Thus

$$P_{H_0}\left(|T_k| > q_{1-\frac{\alpha}{2}}^{(t_{n-p})}\right) = 1 - P\left(-q_{1-\frac{\alpha}{2}}^{(t_{n-p})} \leq t_{n-p} \leq q_{1-\frac{\alpha}{2}}^{(t_{n-p})}\right) = \alpha$$

Test: Reject H_0 if $|T_k| > q_{1-\frac{\alpha}{2}}^{(t_{n-p})}$.

The Gaussian linear model: Statistical inference

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}\sqrt{[(X'X)^{-1}]_k}} \sim t_{n-p}, \qquad \text{Student-t distribution}$$

Thus, with $\hat{se}_k := \hat{\sigma}\sqrt{[(X'X)^{-1}]_k}$,

$$\begin{aligned} &P\left(\hat{\beta}_k - q^{(t_{n-p})}_{1-\frac{\alpha}{2}}\hat{se}_k \leq \beta_k \leq \hat{\beta}_k + q^{(t_{n-p})}_{1-\frac{\alpha}{2}}\hat{se}_k\right) \\ &\quad = P\left(-q^{(t_{n-p})}_{1-\frac{\alpha}{2}} \leq \frac{\hat{\beta}_k - \beta_k}{\hat{se}_k} \leq q^{(t_{n-p})}_{1-\frac{\alpha}{2}}\right) = 1-\alpha \end{aligned}$$

$$\boxed{CI_\alpha = \hat{\beta}_k \pm q^{(t_{n-p})}_{1-\frac{\alpha}{2}}\hat{se}_k}$$

Consider $p = 2$:

$$Y \sim N(X\beta, \sigma^2 I_n), \quad \hat{\beta} = (X'X)^{-1}X'Y \sim N(\beta, \sigma^2(X'X)^{-1})$$

with

$n_1 + n_2 = n$

$$X = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\ \vdots\\1\\1\end{matrix}}\right\} n_1 \\ \left.\vphantom{\begin{matrix}0\\ \vdots\\0\end{matrix}}\right\} n_2 \end{matrix}$$

$$X\beta = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_2 \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} n_1 \\ \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} n_2 \end{matrix}$$

$$Y \sim N(X\beta, \sigma^2 I_n) \Rightarrow \quad \begin{matrix} Y_1 \dots Y_{n_1} \overset{iid}{\sim} N(\beta_1, \sigma^2) \\ Y_{n_1+1} \dots Y_n \overset{iid}{\sim} N(\beta_2, \sigma^2) \end{matrix}$$

$$\hat\beta = (X'X)^{-1} X'Y = \begin{pmatrix} \hat\beta_1 \\ \hat\beta_2 \end{pmatrix} \qquad X'X = \begin{pmatrix} 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 \end{pmatrix} X$$

$$(X'X)^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix} \qquad = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}$$

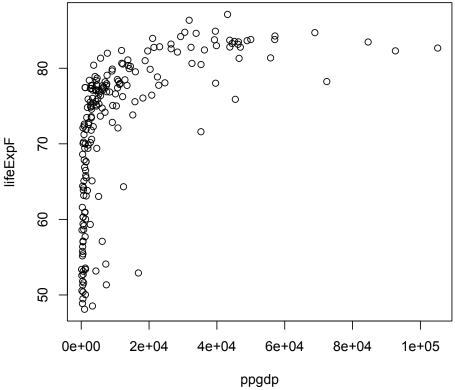
$$X'Y = \begin{pmatrix} 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_{n_1 n_2} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n_1} Y_i \\ \sum_{i=n_1+1}^{n} Y_i \end{pmatrix}$$

$$\hat\beta = \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix} \begin{pmatrix} \sum Y_i \\ \sum Y_i \end{pmatrix} = \begin{pmatrix} \frac{1}{n_1} \sum_{i=1}^{n_1} Y_i \\ \frac{1}{n_2} \sum_{i=n_1+1}^{n} Y_i \end{pmatrix}$$

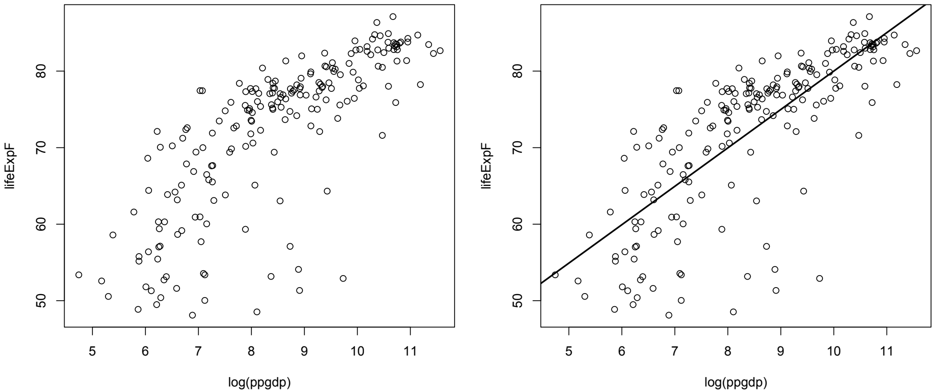
EXAMPLE: UN DATA, 2009

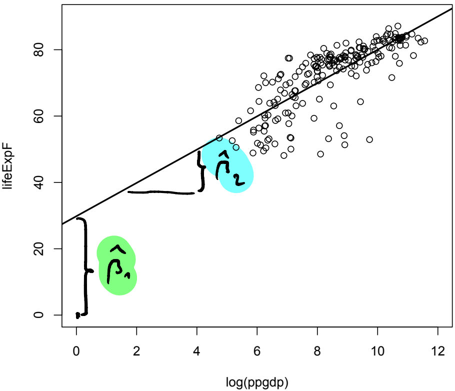
	country	region	group	fertility	ppgdp	lifeExpF	pctUrban
1	Afghanistan	Asia	other	5.968	499.0	49.49	23
2	Albania	Europe	other	1.525	3677.2	80.40	53
3	Algeria	Africa	africa	2.142	4473.0	75.00	67
4	Angola	Africa	africa	5.135	4321.9	53.17	59
5	Anguilla	Caribbean	other	2.000	13750.1	81.10	100
6	Argentina	Latin Amer	other	2.172	9162.1	79.89	93

$n = 199$



Example: UN data





$$X = \begin{pmatrix} 1 & X_{12} \\ 1 & X_{22} \\ \vdots & \vdots \\ 1 & X_{n2} \end{pmatrix}$$

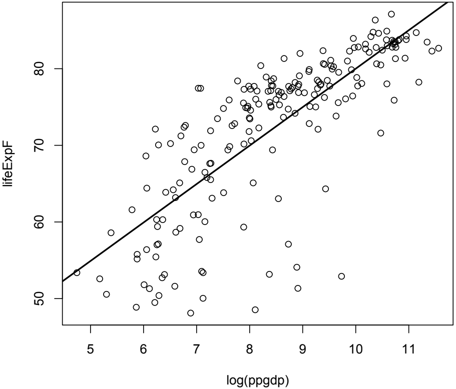
log(ppgdp)

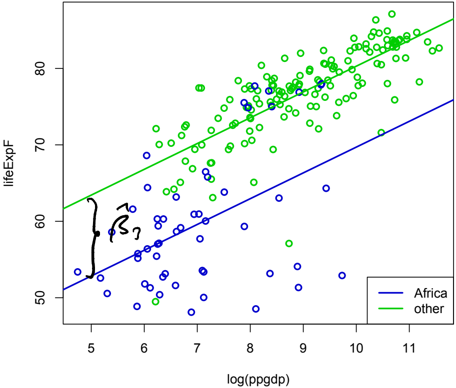
$$E[Y_i] = X_{i\cdot}\beta = \beta_1 + X_{i2}\beta_2$$

Coefficients:

		Estimate	Std. Error	t value	Pr(>|t|)
β_1	(Intercept)	29.8148	2.5314	11.78	<2e-16
β_2	log(ppgdp)	5.0188	0.2942	17.06	<2e-16
		$\hat{\beta}$	$\hat{se}$	T_k	$H_0: \beta_k = 0$

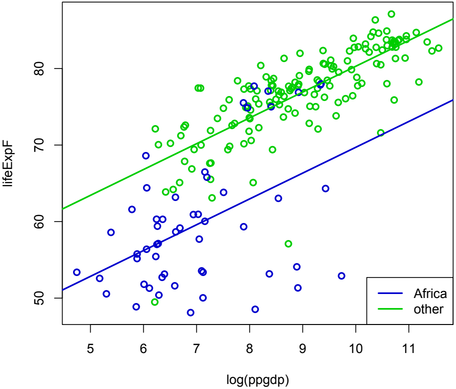
EXAMPLE: UN DATA





		Estimate	Std. Error	t value	Pr(>|t|)
β_1	(Intercept)	35.9798	2.0889	17.22	<2e-16
β_2	log(ppgdp)	3.3728	0.2788	12.10	<2e-16
β_3	groupother	10.5859	0.9802	10.80	<2e-16

Example: UN data



$$X = \begin{pmatrix} 1 & GDP_1 & 1 \\ 1 & GDP_2 & 0 \\ \vdots & \vdots & \\ 1 & GDP_n & 0 \end{pmatrix}$$

group

$$E[Y_i] = X_{i\cdot}\beta = \beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35.9798	2.0889	17.22	<2e-16
log(ppgdp)	3.3728	0.2788	12.10	<2e-16
groupother	10.5859	0.9802	10.80	<2e-16

EXAMPLE: UN DATA

universität wien

group

$$\mathbb{E}[Y_i] = X_{i\cdot}\beta = \beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3$$

$X_{\cdot 3}$ is the 'group' variable where $0 =$ Africa, $1 =$ other. If the i-th country is in Africa we have

$$\mathbb{E}[Y_i] = \beta_1 + X_{i2}\beta_2 + 0,$$

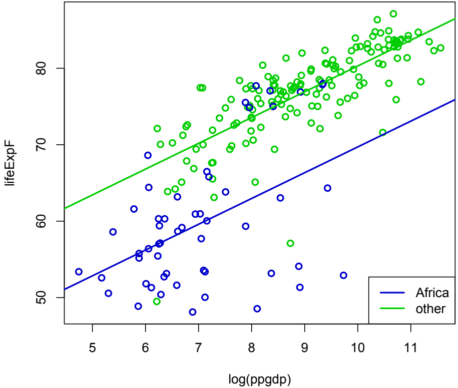
whereas if the j-th country is outside of Africa, we have

$$\mathbb{E}[Y_j] = \beta_1 + X_{j2}\beta_2 + \beta_3.$$

β_3 is the expected additional life expectancy of women in non-African countries, given that $log(ppgdp)$ is the same ($X_{i2} = X_{j2}$).

$$\text{test} \quad H_0 : \beta_3 = 0, \beta_1, \beta_2 \in \mathbb{R}, \sigma^2 > 0$$

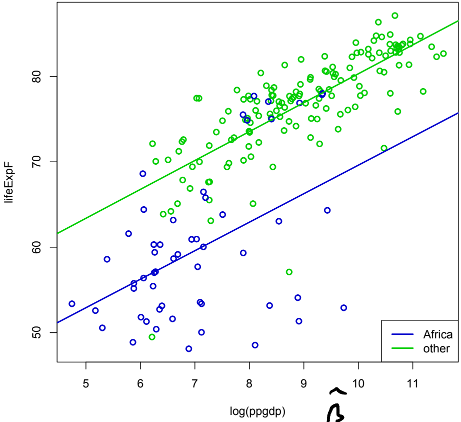
UN DATA: INTERACTION EFFECT



Here we forced the two regression lines to be parallel!

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   35.9798     2.0889   17.22   <2e-16
log(ppgdp)     3.3728     0.2788   12.10   <2e-16
groupother    10.5859     0.9802   10.80   <2e-16
```

UN DATA: INTERACTION EFFECT



$$X = \begin{pmatrix} 1 & GDP_1 & 0 & 0 \\ 1 & GDP_2 & 1 & GDP_2 \\ 1 & GDP_3 & 0 & 0 \\ \vdots & & & \\ 1 & GDP_n & 1 & GDP_n \end{pmatrix}$$

group

Africa:

$\mathbb{E}[Y_i] = \beta_1 + X_{i2}\beta_2$

other:

$\mathbb{E}[Y_j] = \beta_1 + X_{j2}\beta_2 + \beta_3 + X_{j2}\beta_4 = \beta_1 + \beta_3 + X_{j2}(\beta_2 + \beta_4)$

$\hat{\beta}$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.22882	4.18145	8.664	1.73e-15
log(ppgdp)	3.33752	0.58428	5.712	4.12e-08
groupother	10.24281	5.08235	2.015	0.0452
log(ppgdp):groupother	0.04578	0.66539	0.069	0.9452

Test: Are the slopes different? $H_0: \beta_4 = 0$