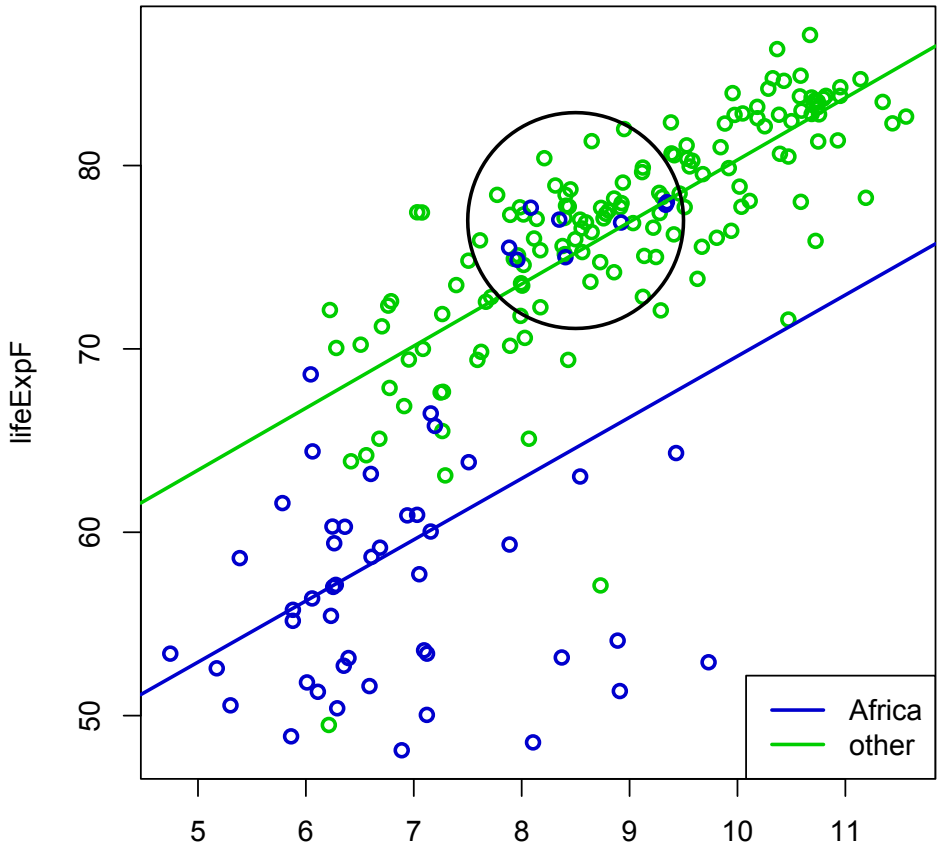
UN DATA: INTERACTION EFFECT

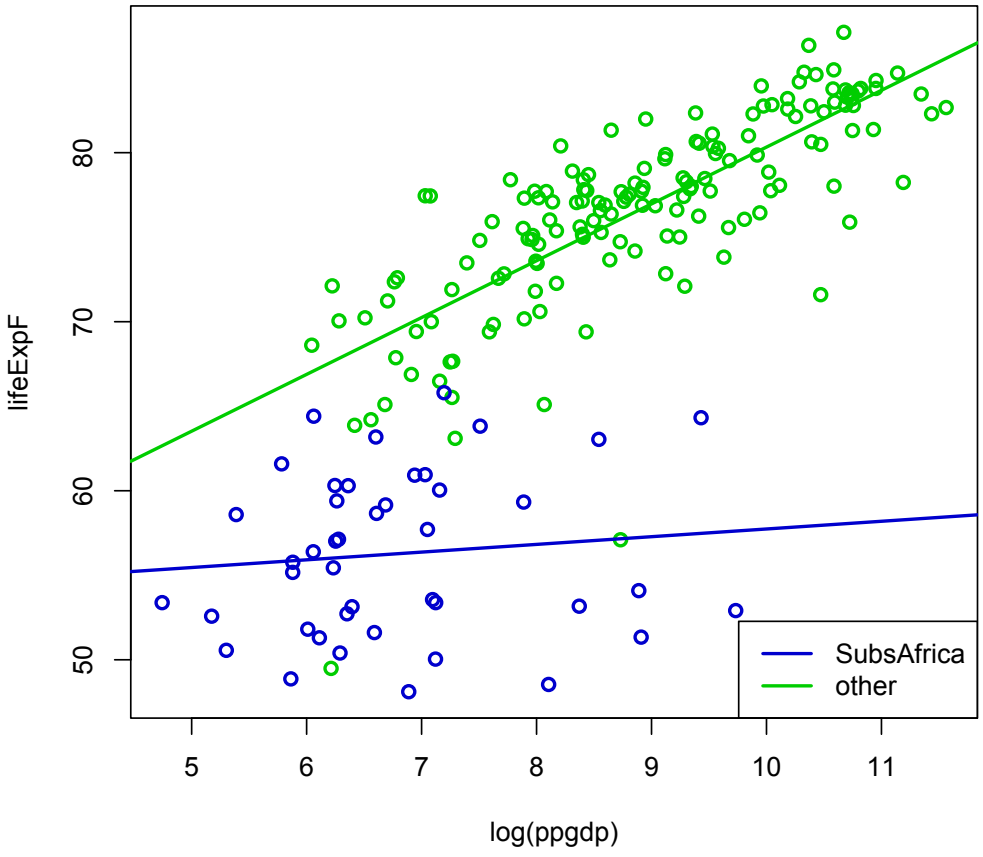


Algeria, Cape Verde, Egypt, Libya, Mauritius, Morocco, Seychelles, Tunisia

North African countries or islands!

	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

UN DATA: INTERACTION EFFECT



sub-Saharan Africa:

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

other:

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

Test: Are the slopes different?

$H_0: \beta_4 = 0$

Does GDP actually influence female life expectancy in sub-Saharan Africa?

		Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	β_1	53.4042	3.9035	13.681	< 2e-16
log(ppgdp)	β_2	0.4330	0.5656	0.766	0.4448
groupSSother	β_3	-7.9811	4.4871	-1.779	0.0769
log(ppgdp):groupSSother	β_4	3.0627	0.6167	4.967	1.48e-06

$\alpha = 0.05$ $\alpha = 0.01$

Looking at the data and then changing the model or the question to ask is actually not allowed!!

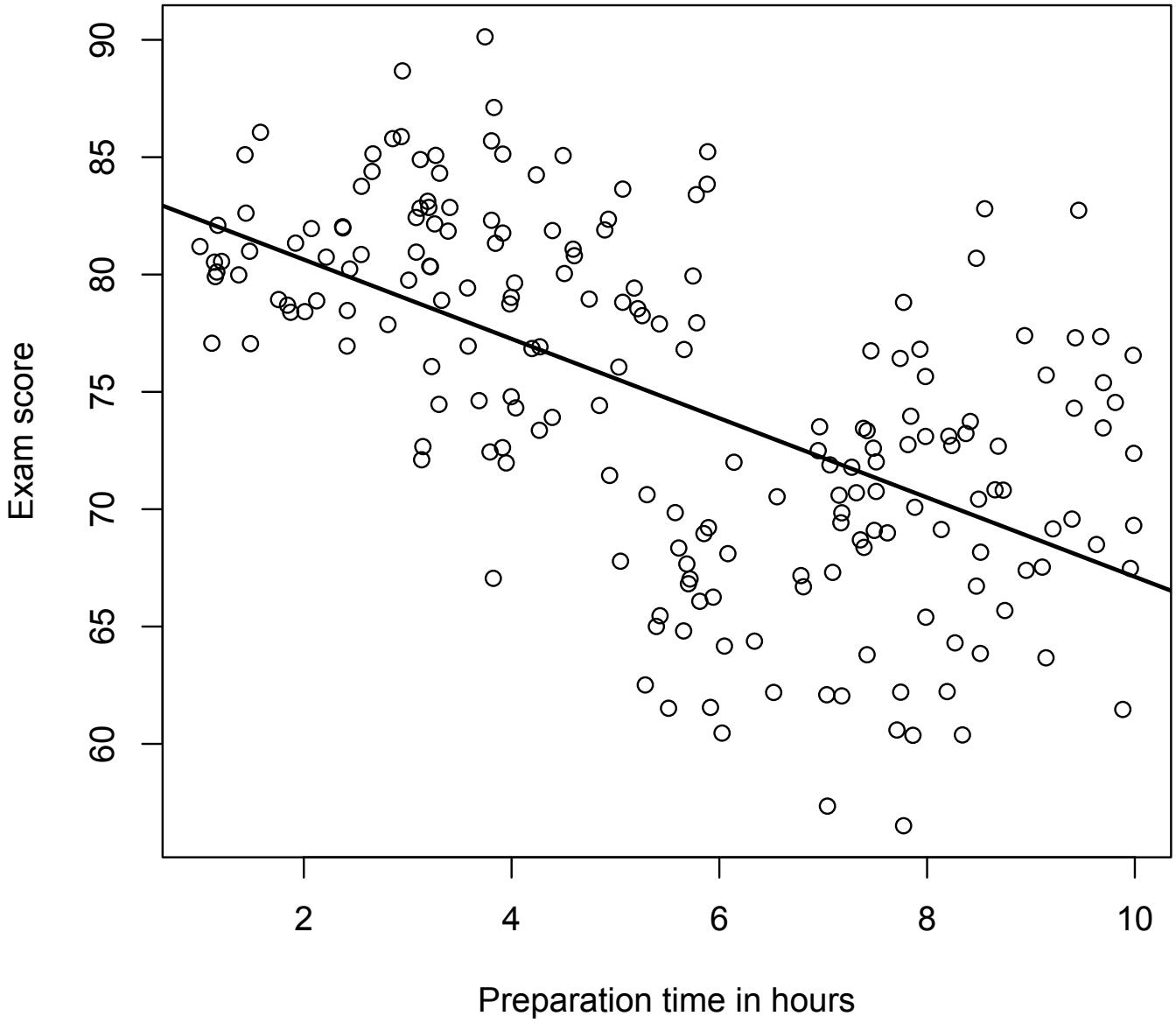
More on that later...

Exploratory data analysis vs. statistical inference!

Philosophical question: What actually is the population here?

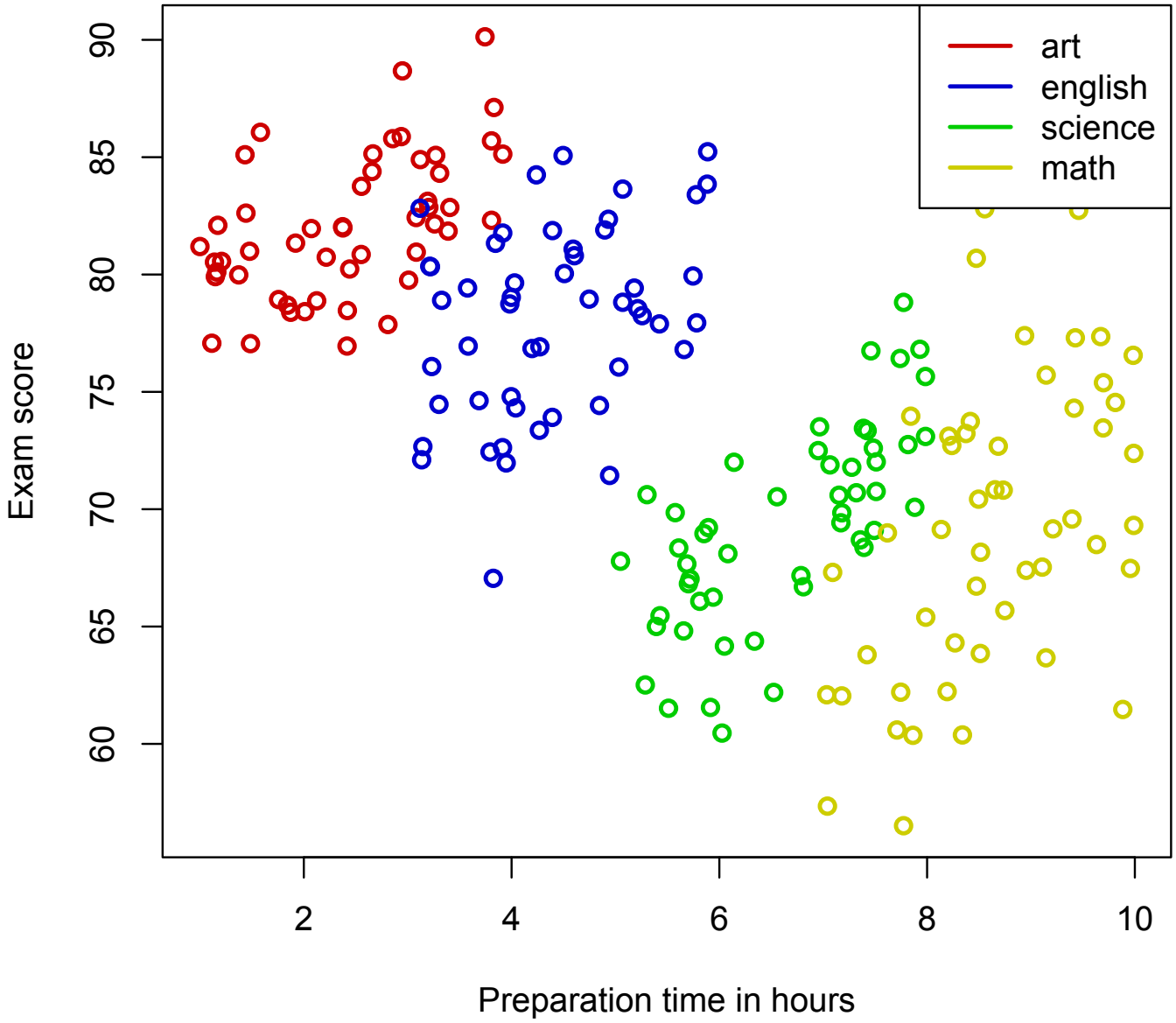
SIMPSON'S PARADOX

universität wien



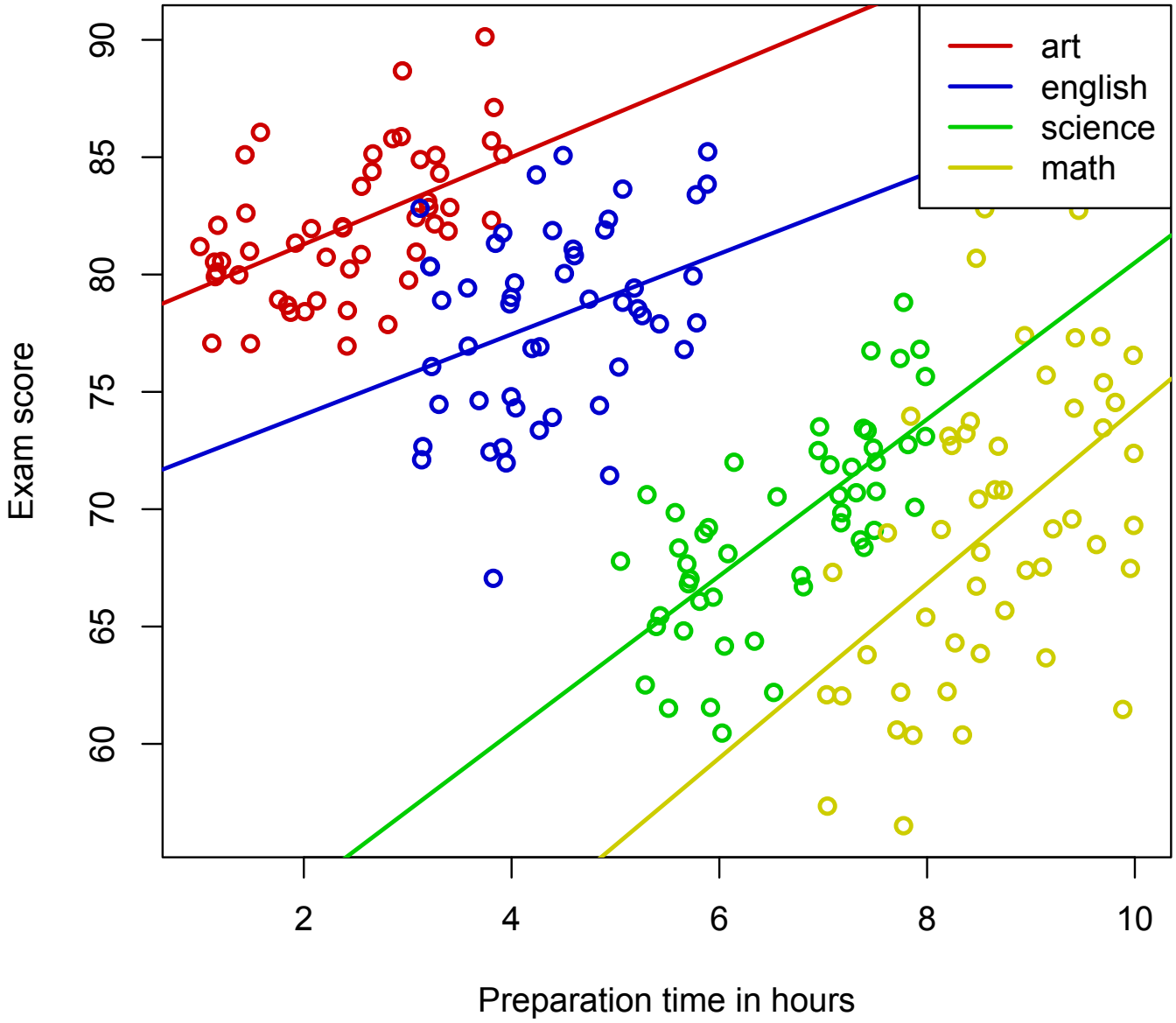
Simpson's paradox

universität wien

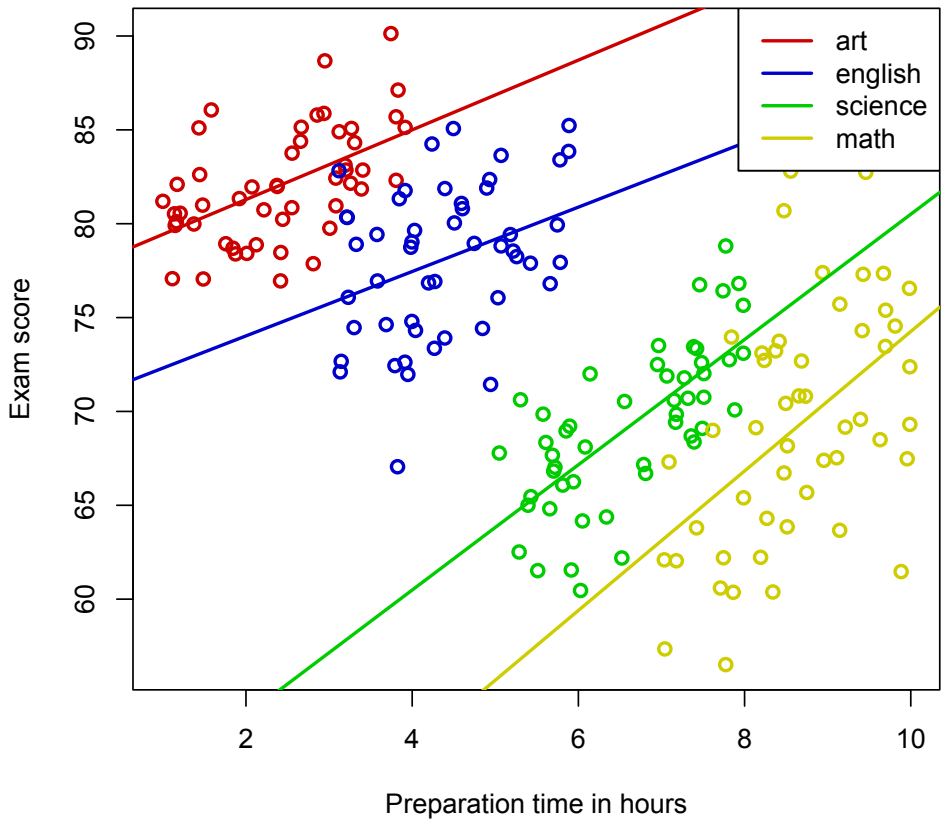


Simpson's paradox

universität wien



Categorical variables aka factors



Are the four slopes actually different?

Here 'subject' is a categorical variable (*factor*) with four *levels*. How to model that, i.e., how to construct X?

Categorical variables aka factors

How to model a factor variable?

1.) Use a code: art=1, english=2, science=3, math=4.

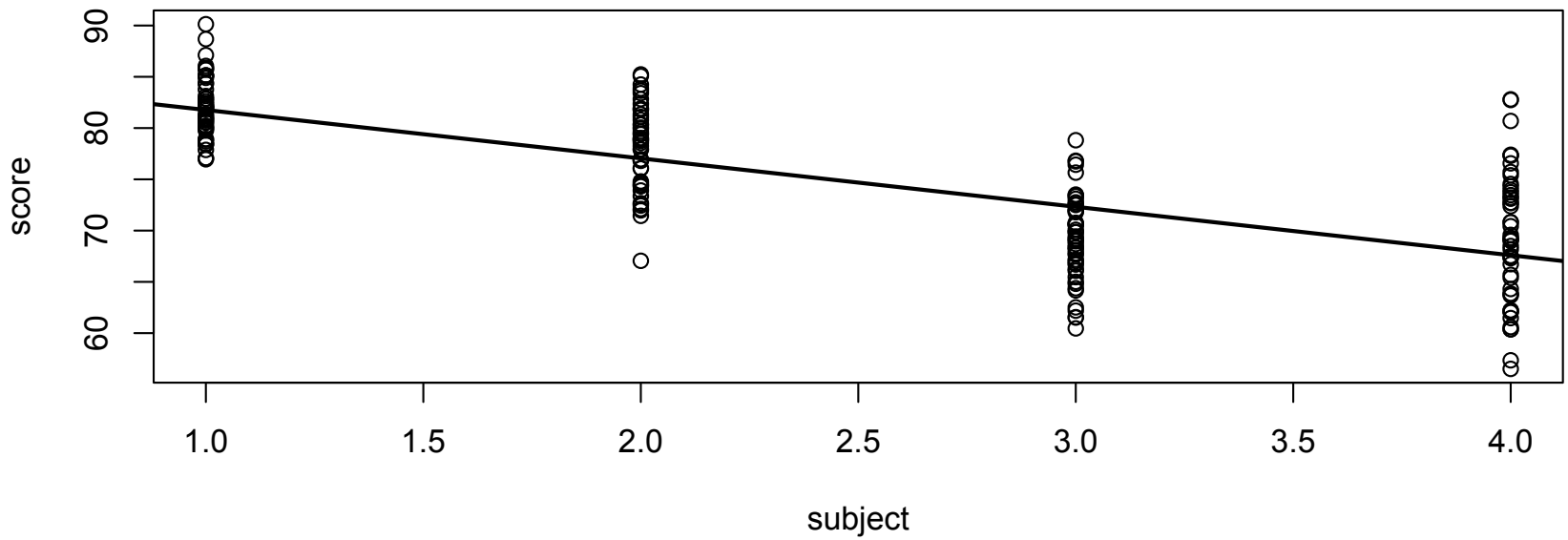
$$X = \begin{pmatrix} 1 & 2.5 & 1 \\ 1 & 3.5 & 2 \\ 1 & 6 & 4 \\ 1 & 7.5 & 3 \\ 1 & 5 & 3 \\ 1 & 1 & 1 \\ 1 & 1.5 & 2 \\ & \vdots & \\ 1 & 9 & 3 \end{pmatrix}$$

time

subject

Categorical variables aka factors

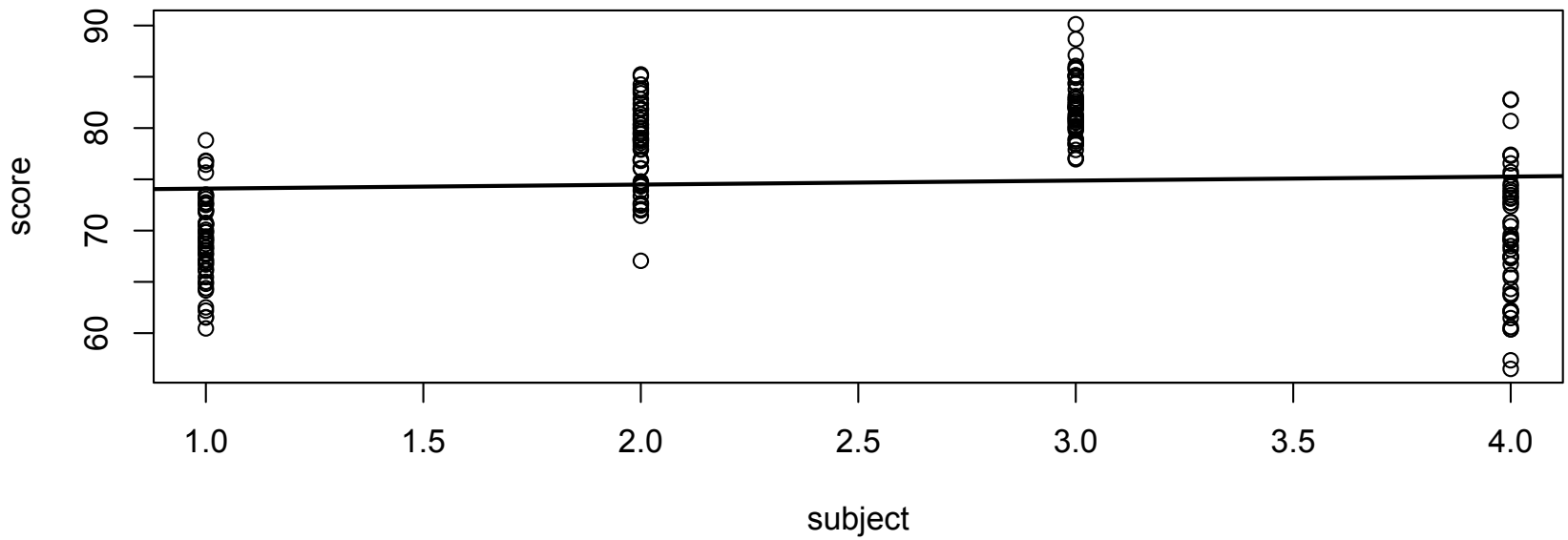
1a.) Use a code: art=1, english=2, science=3, math=4.




```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  86.4877     0.8512   101.6   <2e-16
subj         -4.7225     0.3108   -15.2   <2e-16
```

Categorical variables aka factors

1b.) Use a different code: art=3, english=2, science=1, math=4.




```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  73.7040     1.2504  58.946   <2e-16
subj          0.3911     0.4566   0.856    0.393
```

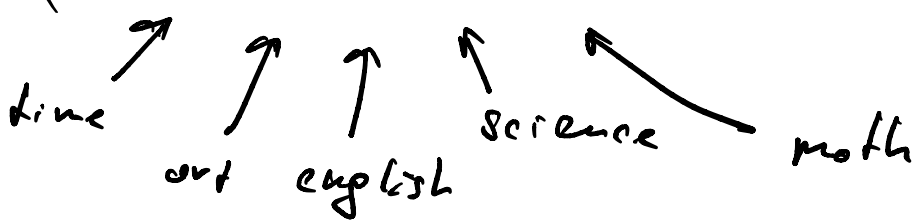
CATEGORICAL VARIABLES AKA FACTORS

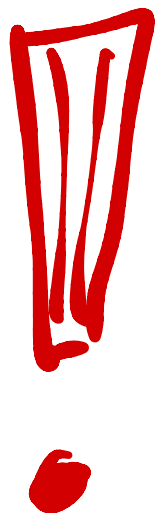
universität wien

2.) Introduce 'dummy variables'.

$X_{\cdot 7} + X_{\cdot 4} + X_{\cdot 5} + X_{\cdot 6} = X_{\cdot 1}$

$$X = \begin{pmatrix} 1 & 2.5 & 1 & 0 & 0 & 0 & 2.5 & 0 & 0 & 0 \\ 1 & 3.5 & 0 & 1 & 0 & 0 & 0 & 3.5 & 0 & 0 \\ 1 & 6 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 6 \\ 1 & 7.5 & 0 & 0 & 1 & 0 & 0 & 0 & 7.5 & 0 \\ 1 & 5 & 0 & 0 & 1 & 0 & 0 & 0 & 5 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1.5 & 0 & 1 & 0 & 0 & 0 & 1.5 & 0 & 0 \\ & \vdots & & & & & & \vdots & & \\ 1 & 9 & 0 & 0 & 1 & 0 & 0 & 0 & 9 & 0 \end{pmatrix}$$





Test if the slopes are all the same, i.e. $H_0 : \beta_7 = \beta_8 = \beta_9 = \beta_{10}$

Categorical variables aka factors

3a.) Avoiding the *dummy variable trap*:

$$X = \begin{pmatrix} 1 & 2.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 3.5 & 1 & 0 & 0 & 3.5 & 0 & 0 \\ 1 & 6 & 0 & 0 & 1 & 0 & 0 & 6 \\ 1 & 7.5 & 0 & 1 & 0 & 0 & 7.5 & 0 \\ 1 & 5 & 0 & 1 & 0 & 0 & 5 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1.5 & 1 & 0 & 0 & 1.5 & 0 & 0 \\ & \vdots & & & & & \vdots & \\ 1 & 9 & 0 & 1 & 0 & 0 & 9 & 0 \end{pmatrix}$$

english, science, math

$$\text{all}: \mathbb{E}(Y_i) = \beta_1 + X_{i2}\beta_2$$

$$\text{engl.}: \mathbb{E}(Y_i) = \beta_1 + X_{i2}\beta_2 + \beta_3 + X_{i2}\beta_6 = \beta_1 + \beta_3 + X_{i2}(\beta_2 + \beta_6)$$

$$\text{math}: \mathbb{E}(Y_i) = \beta_1 + X_{i2}\beta_2 + \beta_5 + X_{i2}\beta_8 = \beta_1 + \beta_5 + X_{i2}(\beta_2 + \beta_8)$$

Test if the slopes are all the same, i.e., $H_0 : \beta_6 = \beta_7 = \beta_8 = 0$.

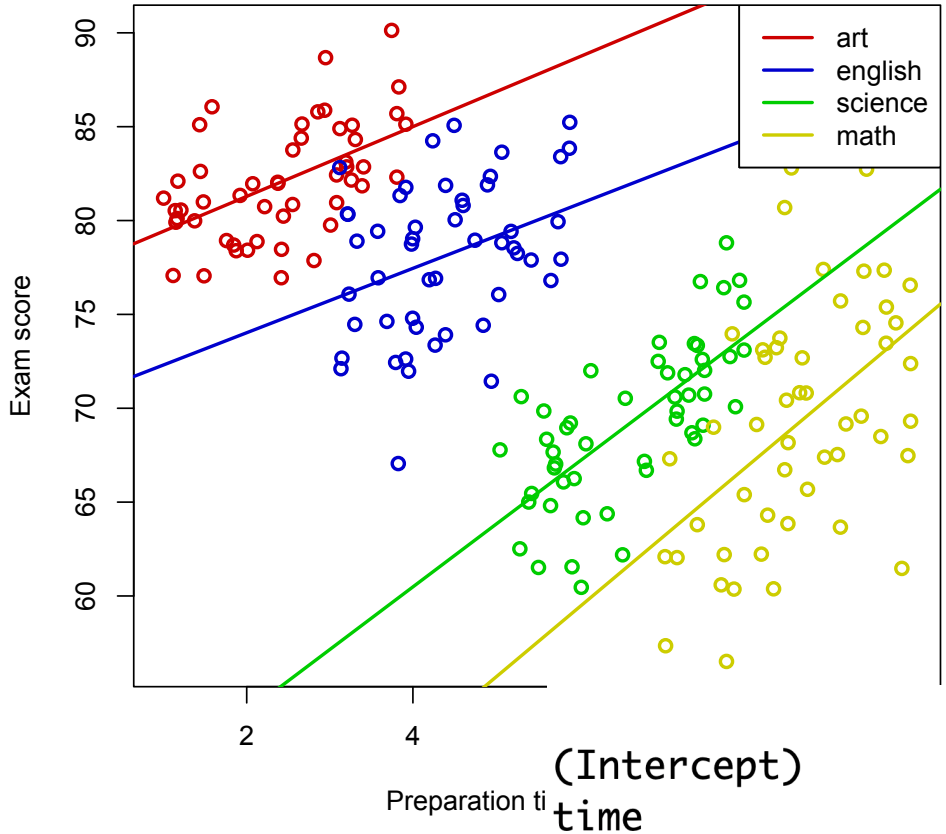
Categorical variables aka factors

3b.) Avoiding the *dummy variable trap*:

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 & 2.5 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 3.5 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 6 \\ 0 & 0 & 1 & 0 & 0 & 0 & 7.5 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 5 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1.5 & 0 & 0 \\ & & \vdots & & & \vdots & & \\ 0 & 0 & 1 & 0 & 0 & 0 & 9 & 0 \end{pmatrix}$$

Test if the slopes are all the same, i.e., $H_0 : \beta_5 = \beta_6 = \beta_7 = \beta_8$.

Categorical variables aka factors



Are the four slopes actually different?

$H_0: \beta_6 = \beta_7 = \beta_8 = 0$

$\alpha = 0.05$

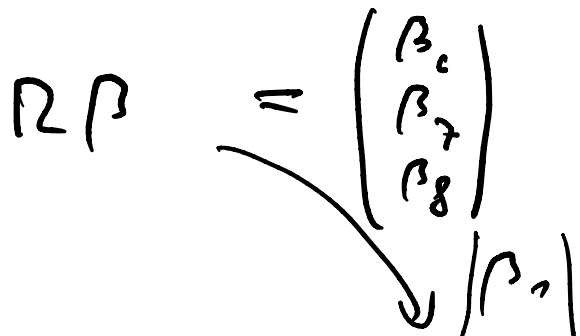
		Estimate	Std. Error	t value	Pr(>|t|)
β_1	(Intercept)	77.5733	1.6652	46.585	< 2e-16
β_2	time	1.8569	0.6487	2.863	0.00467
	subjectenglish	-6.9800	3.4766	-2.008	0.04608
	subjectmath	-40.4671	5.9648	-6.784	1.41e-10
	subjectscience	-30.4375	4.5206	-6.733	1.87e-10
β_6	time:subjectenglish	-0.1419	0.9432	-0.150	0.88053
β_7	time:subjectmath	1.8583	0.9243	2.010	0.04579
β_8	time:subjectscience	1.4801	0.9026	1.640	0.10270

Can we test this only by looking at the standard output? no!

We want to test a general linear hypothesis:

$$H_0 : R\beta = r$$

where $R \in \mathbb{R}^{q \times p}$, $q \leq p$, $\operatorname{rank} R = q$ and $r \in \mathbb{R}^q$.



For instance, with

$$R = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad r = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

we have

$$H_0 : R\beta = r \quad \Longleftrightarrow \quad H_0 : \beta_6 = 0, \beta_7 = 0, \beta_8 = 0$$

We want to test a general linear hypothesis:

$$H_0 : R\beta = r$$

where $R \in \mathbb{R}^{q \times p}$, $q \leq p$, $\operatorname{rank} R = q$ and $r \in \mathbb{R}^q$.

Or, with $R\beta = \begin{pmatrix} \beta_5 - \beta_6 \\ \beta_6 - \beta_7 \\ \beta_7 - \beta_8 \end{pmatrix}$

$$R = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \quad \text{and} \quad r = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

we have

$$H_0 : R\beta = r \quad \Longleftrightarrow \quad H_0 : \beta_5 = \beta_6, \beta_6 = \beta_7, \beta_7 = \beta_8$$

$\beta_5 = \beta_6 = \beta_7 = \beta_8$

$$H_0 : R\beta = r$$

Notice:

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}) \quad \Rightarrow \quad R\hat{\beta} - r \sim N(R\beta - r, \sigma^2 R(X'X)^{-1}R')$$

$R\beta - r = 0$ under H_0

Under H_0, we therefore have

$$\frac{[R(X'X)^{-1}R']^{-1/2}(R\hat{\beta} - r)}{\sigma} \sim N(0, I_q)$$

and

$$\frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{\sigma^2} \sim \chi^2_q.$$

$$H_0 : R\beta = r$$

$$\frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{\sigma^2} \sim \chi^2_q.$$

Recall, $\hat{\sigma}^2 := \frac{1}{n-p}\|Y - X\hat{\beta}\|_2^2$ satisfies $\frac{\hat{\sigma}^2(n-p)}{\sigma^2} \sim \chi^2_{n-p}$ independent of $\hat{\beta}_n$.

Definition:
If $S_1 \sim \chi^2_{d_1}$ independent of $S_2 \sim \chi^2_{d_2}$, then $\frac{S_1/d_1}{S_2/d_2} \sim F_{d_1,d_2}$ follows an F-distribution with d_1 and d_2 degrees of freedom.

Hence, under H_0,

$$F := \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/q}{\hat{\sigma}^2} \sim F_{q,n-p}$$

$$H_0 : R\beta = r$$

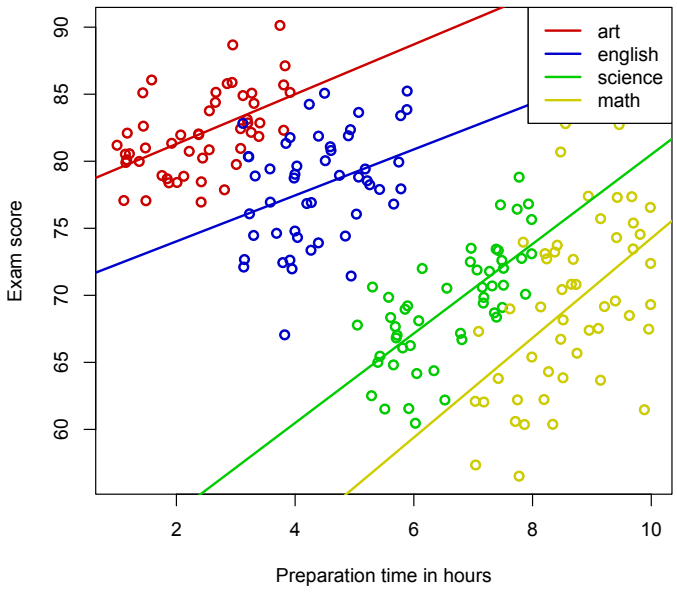
Under H_0,

$$F := \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/q}{\hat{\sigma}^2} \sim F_{q,n-p}$$

Test: Reject H_0 if $F > c_\alpha := q_{1-\alpha}^{(F_{q,n-p})}$.

Hence, $P_{H_0}(F > c_\alpha) = 1 - P_{H_0}(F \leq c_\alpha) = 1 - (1 - \alpha) = \alpha$.

Testing general linear hypotheses



Are the four slopes actually different?

$$H_0 : \beta_6 = \beta_7 = \beta_8 = 0 \iff H_0 : R\beta = r$$

$$R = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, r = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$\alpha = 0.05$

q $\quad$ $n-p$

```
F = 2.3877, df1 =   3, df2 = 192, p-value = 0.07029
```

Is there a contradiction with the marginal t-tests?

What if we used another R matrix?

HW

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	77.5733	1.6652	46.585	< 2e-16
time	1.8569	0.6487	2.863	0.00467
subjectenglish	-6.9800	3.4766	-2.008	0.04608
subjectmath	-40.4671	5.9648	-6.784	1.41e-10
subjectscience	-30.4375	4.5206	-6.733	1.87e-10
time:subjectenglish	-0.1419	0.9432	-0.150	0.88053
time:subjectmath	1.8583	0.9243	2.010	0.04579
time:subjectscience	1.4801	0.9026	1.640	0.10270

$H_0: \beta_6 = 0$

Test $H_0 : \beta_6 = \beta_7 = \beta_8 = 0$ at level $\alpha = 0.05$.

```
F = 2.3877, df1 =  3, df2 = 192, p-value = 0.07029
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	77.5733	1.6652	46.585	< 2e-16
time	1.8569	0.6487	2.863	0.00467
subjectenglish	-6.9800	3.4766	-2.008	0.04608
subjectmath	-40.4671	5.9648	-6.784	1.41e-10
subjectscience	-30.4375	4.5206	-6.733	1.87e-10
time:subjectenglish	-0.1419	0.9432	-0.150	0.88053
time:subjectmath	1.8583	0.9243	2.010	0.04579
time:subjectscience	1.4801	0.9026	1.640	0.10270

$\leq \alpha$?

$\leq \frac{\alpha}{3}$?

Is there a contradiction with the marginal t-tests?

$\alpha = 0.05 \rightarrow \frac{\alpha}{3} = 0.0166$

F-TEST VS. MARGINAL T-TESTS

$H_0: \beta_6 = \beta_7 = \beta_8 = 0$

$$\varphi_\alpha(Y) := \begin{cases} 1, & \text{if } \varphi_\alpha^{(k)}(Y) = 1 \text{ for some } k \in \{6,7,8\} \\ 0, & \text{else} \end{cases}$$

where $\varphi_\alpha^{(k)}(Y) = \begin{cases} 1, & \text{if } |T_k(Y)| > q_{1-\frac{\alpha}{2}}^{(t_{n-p})} \\ 0, & \text{else} \end{cases}$

is the t-test for $H_0^{(k)}: \beta_k = 0$ which satisfies

$$P_{H_0^{(k)}}(\varphi_\alpha^{(k)} = 1) = \alpha.$$

Is φ_α a level-α test for H_0?

$$P_{H_0}(\varphi_\alpha = 1) \leq P_{H_0}\Big(\varphi_\alpha^{(6)} = 1 \cup \varphi_\alpha^{(7)} = 1 \cup \varphi_\alpha^{(8)} = 1\Big)$$

$$\leq \sum_{k=6}^{8} \underbrace{P_{H_0}(\varphi_\alpha^{(k)} = 1)}_{= \alpha \text{ because } H_0 \subseteq H_0^{(k)}} = 3 \cdot \alpha$$

For many X matrices we can have

$$P_{H_0}(\varphi_\alpha = 1) > \alpha .$$

note: $P_{H_0}(\varphi_{\alpha/3} = 1) \leq 3 \cdot \frac{\alpha}{3} = \alpha$

universität wien

Theorem (Bonferroni correction)

Let $(\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$ be a statistical model, $\alpha \in (0,1)$ and $\Theta_0^{(1)}, \ldots, \Theta_0^{(K)} \subseteq \Theta$ be hypotheses with corresponding level-α tests $\varphi_\alpha^{(1)}, \ldots \varphi_\alpha^{(K)}$, i.e., for every $k \in \{1, \ldots, K\}$, we have $P_\theta(\varphi_\alpha^{(k)} = 1) \leq \alpha$ for all $\theta \in \Theta_0^{(k)}$. Then the test

$$\varphi_\alpha(x) := \begin{cases} 1, & \text{if } \exists k \in \{1, \ldots, K\} : \varphi_\alpha^{(k)}(x) = 1, \\ 0, & \text{else,} \end{cases}$$

satisfies

$$P_\theta(\varphi_{\frac{\alpha}{K}} = 1) \leq \alpha \quad \forall \theta \in \Theta_0 := \bigcap_{k=1}^{K} \Theta_0^{(k)}.$$

Multiple testing

(Intercept)	2.299346	1.307025	1.759	0.08010
Xpopulation	-3.477014	2.323128	-1.497	0.13608
Xhouseholdsize	1.118523	0.598059	1.870	0.06294
Xracepctblack	-0.042317	0.205327	-0.206	0.83693
XracePctWhite	0.022103	0.290556	0.076	0.93944
XracePctAsian	-0.084169	0.139117	-0.605	0.54587
XracePctHisp	0.151692	0.204243	0.743	0.45855
XagePct12t21	-0.239641	0.598155	-0.401	0.68913
XagePct12t29	-0.870986	0.773690	-1.126	0.26165
XagePct16t24	0.381891	0.895346	0.427	0.67019
XagePct65up	0.199093	0.585702	0.340	0.73428
XnumbUrban	3.802921	2.348860	1.619	0.10704
XpctUrban	-0.199798	0.167798	-1.191	0.23521
XmedIncome	-0.386196	0.864797	-0.447	0.65568
XpctWWage	-0.596946	0.551386	-1.083	0.28030
XpctWFarmSelf	-0.308686	0.148121	-2.084	0.03846
XpctWInvInc	-0.751876	0.291920	-2.576	0.01074
XpctWSocSec	-0.721007	0.574259	-1.256	0.21078
XpctWPubAsst	0.238305	0.214464	1.111	0.26786
XpctWRetire	0.064647	0.194944	0.332	0.74053
XmedFamInc	0.871977	0.641222	1.360	0.17543
XperCapInc	-0.876057	0.690411	-1.269	0.20598
XwhitePerCap	0.355919	0.418602	0.850	0.39622
XblackPerCap	-0.302640	0.226430	-1.337	0.18291
XindianPerCap	0.016152	0.120405	0.134	0.89342
XAsianPerCap	-0.147430	0.123883	-1.190	0.23546
XOtherPerCap	0.137916	0.128604	1.072	0.28486
XHispPerCap	-0.072551	0.146598	-0.495	0.62123
XNumUnderPov	-0.392185	0.314491	-1.247	0.21387
XPctPopUnderPov	0.525168	0.363595	1.444	0.15023
XPctLess9thGrade	-0.110777	0.317642	-0.349	0.72765
XPctNotHSGrad	-0.696948	0.439727	-1.585	0.11459
XPctBSorMore	-0.376918	0.385047	-0.979	0.32884

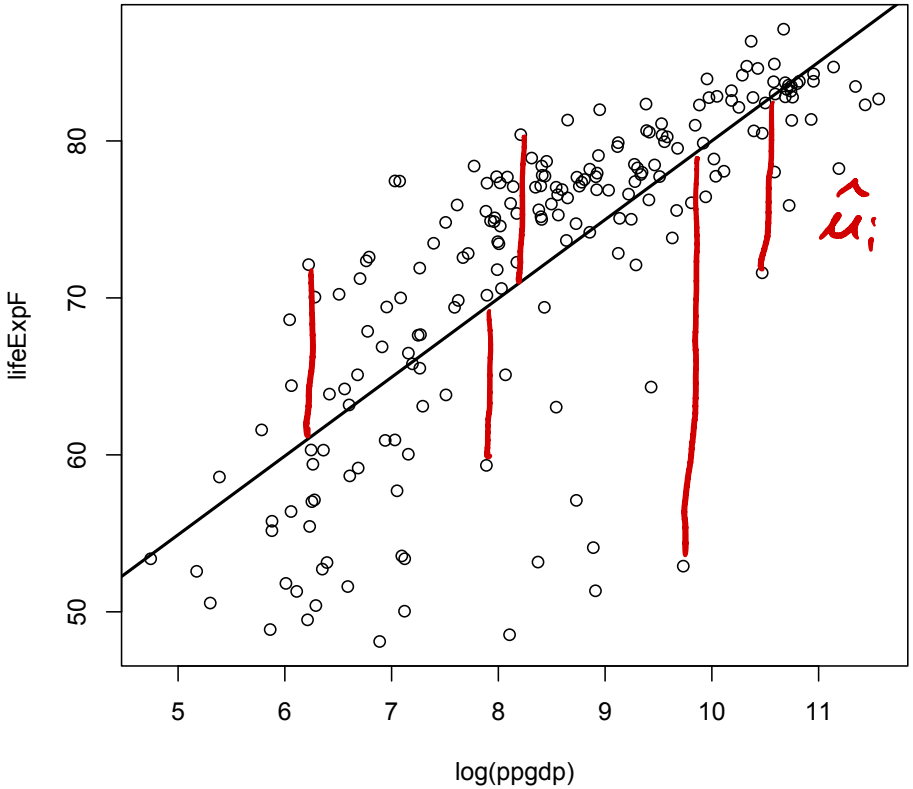
Looking at an output like that and asking, *is there something significant*, is not a good idea!

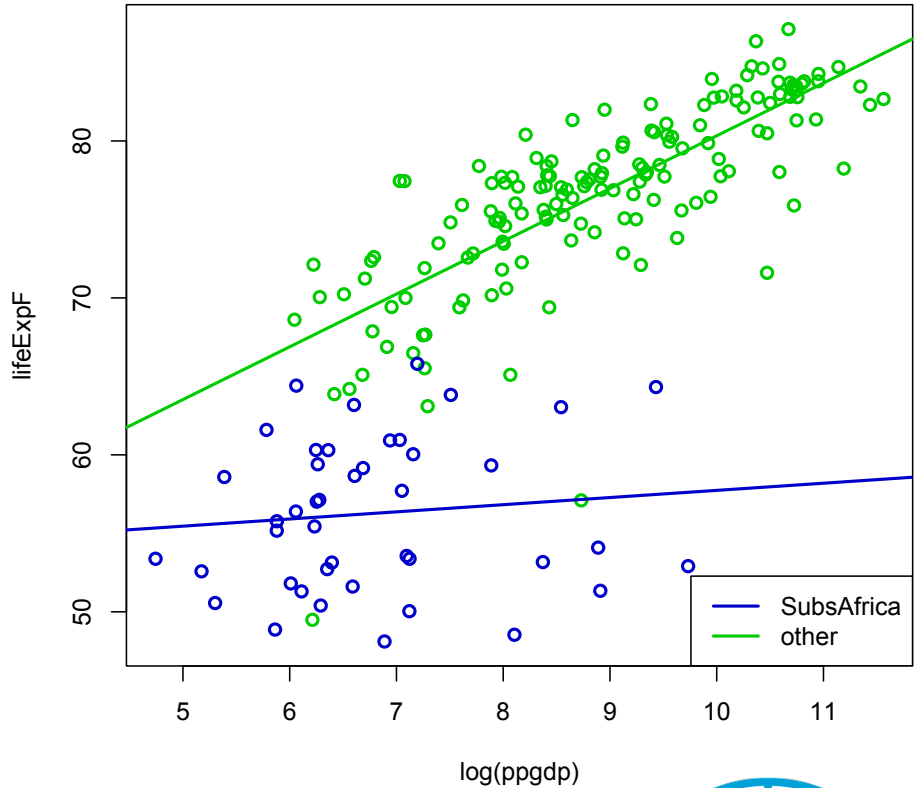
$$H_0 : \forall k : \beta_k = 0$$
$$H_1 : \exists k : \beta_k \neq 0$$

Whats the probability of finding something even though there is nothing?

Model diagnostic plots

Gaussian linear Model: $Y \sim N(X\beta, \sigma^2 I_n)$





What if the model is too complex to visualize it like that? Can we still get a graphical representation of model fit?

Model diagnostic plots

Gaussian linear Model:

$$Y \sim N(X\beta, \sigma^2 I_n)$$

or, equivalently

$$Y_i = X_{i\cdot}\beta + u_i, \quad \text{with} \quad u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \sim N(0, \sigma^2 I_n)$$

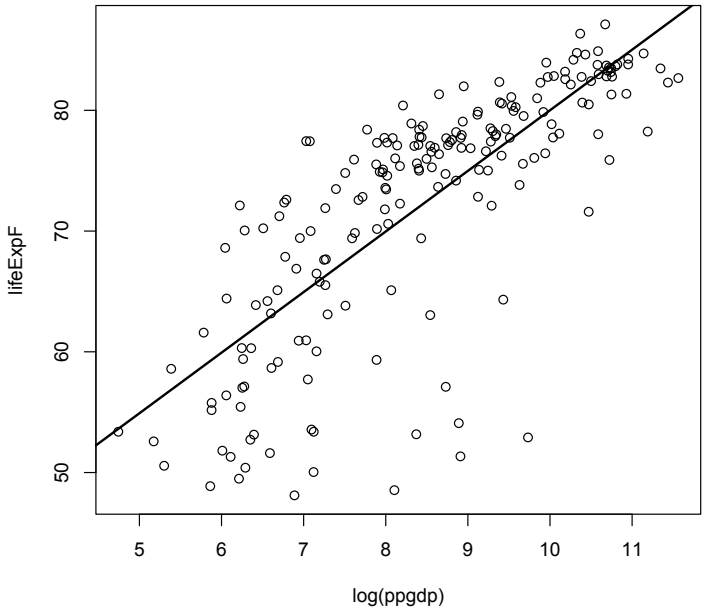
Recall: OLS
$\hat{\beta} = \text{argmin}_{b\in\mathbb{R}^p} \|Y - Xb\|_2^2 = \text{argmin}_{b\in\mathbb{R}^p} \sum_{i=1}^n (Y_i - X_{i\cdot}b)^2$

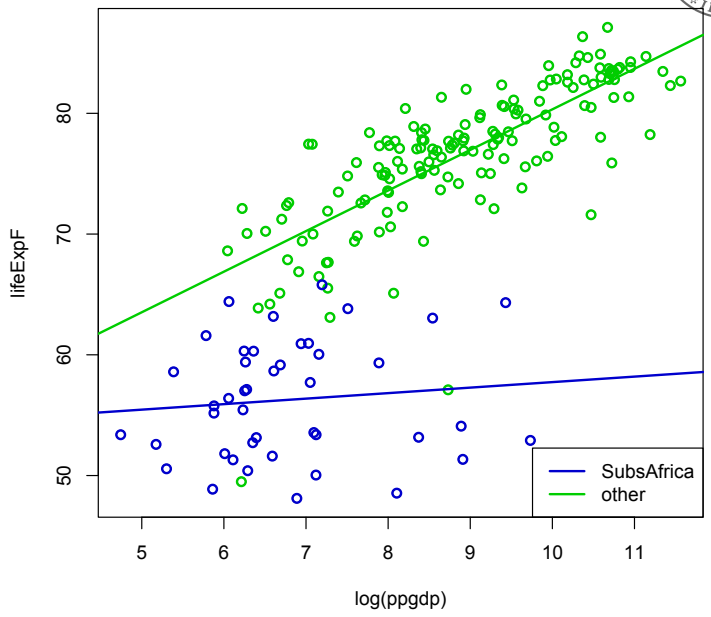
Idea: The *residuals* $\hat{u}_i = Y_i - X_{i\cdot}\hat{\beta}$ should be (approximately) iid Gaussian and evenly distributed around the regression line (independent of their 'location').

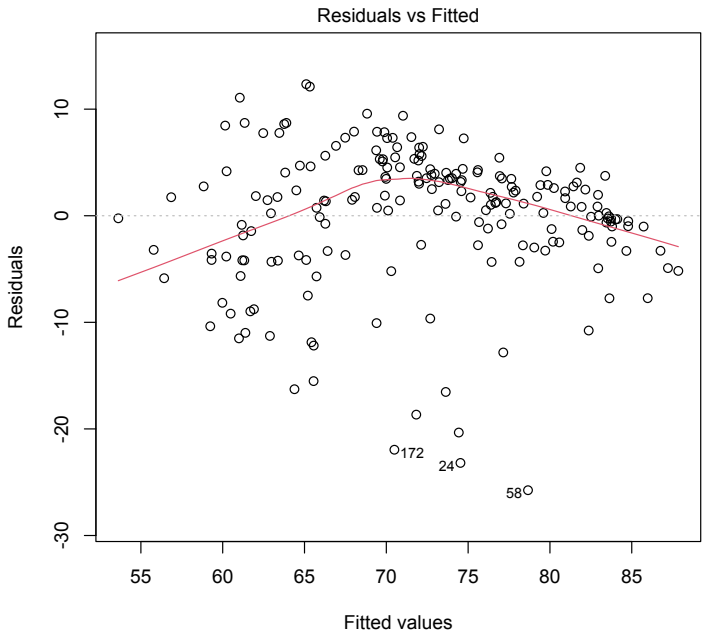
$\hat{Y}_i = X_{i\cdot}\hat{\beta}$ are the *fitted* or *predicted* values 'on' the regression line.

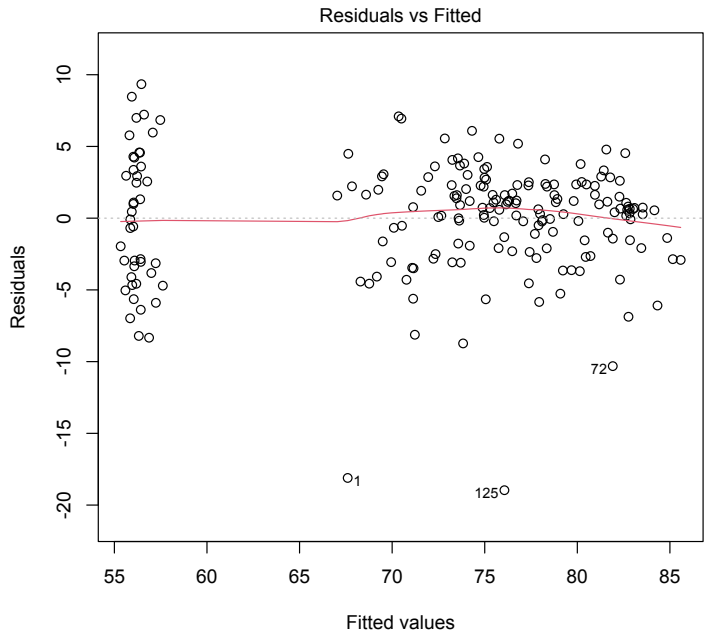
1.) plot $\hat{Y}_i$ against $\hat{u}_i$ 2.) check $\hat{u}_i$ for Gaussianity

RESIDUAL PLOT









1 = Afghanistan
72 = Greenland
125 = Nauru