Correlation and Regression

Vera Demberg

Saarland University

November 30th, 2021

Recap from last week

Significance tests we have looked at so far:

- one or two categorial variables
  $\rightarrow \chi^2$ test / McNemar's test / Fisher's exact test
  (e.g. skin color and death sentence example)

Recap from last week

Significance tests we have looked at so far:

- one or two categorial variables
  $\rightarrow \chi^2$ test / McNemar's test / Fisher's exact test
  (e.g. skin color and death sentence example)
- the response variable (= dependent variable) is ratio scale or interval scale, and the independent variable is categorial
  $\rightarrow$ *t*-test / Welch test / Wilcoxon test
  (e.g. does the student group from one tutor receive higher grades than the student group that's with the other tutor?)

Recap from last week

Significance tests we have looked at so far:

- one or two categorial variables
  $\rightarrow \chi^2$ test / McNemar's test / Fisher's exact test
  (e.g. skin color and death sentence example)
- the response variable ($=$ dependent variable) is ratio scale or interval scale, and the independent variable is categorial
  $\rightarrow$ *t*-test / Welch test / Wilcoxon test
  (e.g. does the student group from one tutor receive higher grades than the student group that's with the other tutor?)
- But what if both the response variable ($=$ dependent variable) and the independent variable is ratio or interval scale?
  $\rightarrow$ today: correlation / regression

Descriptive Statistics vs. Hypothesis Testing

Remind yourself of the difference between "descriptive statistics" and "hypothesis testing".

Let's assume we are wondering about the relationship between *amount of sleep* and *performance*.

- What would you get out of "descriptive statistics"?
- What would you ask if you were doing hypothesis testing?

Table of Contents

Table of Contents

Variance

We have already seen the concept of **variance**.
Take a moment: do you remember what variance is and how the variance of a sample is defined (mathematically)?
Please write it down for yourself.

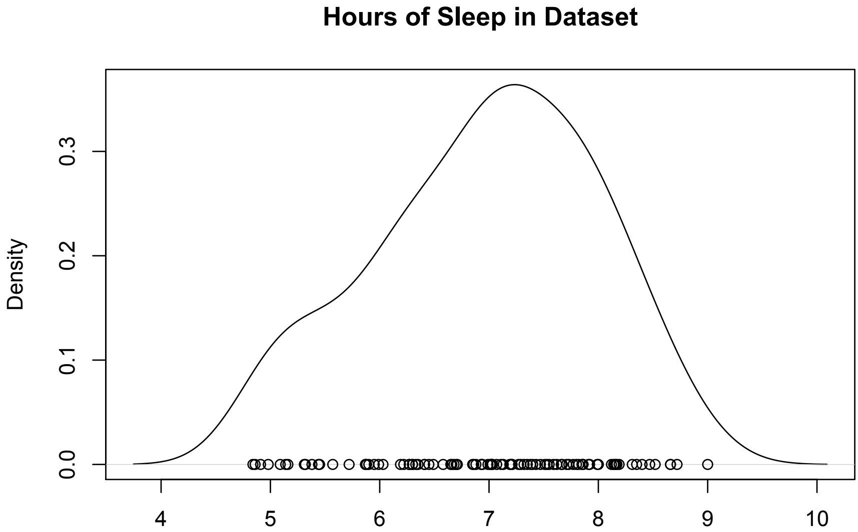
Variance

We have already seen the concept of **variance**.
Take a moment: do you remember what variance is and how the variance of a sample is defined (mathematically)?
Please write it down for yourself.

$$s^2 = \frac{\sum (X - \bar{X})^2}{N - 1}$$

This is the variance of a single sample, which we have previously visualized using a density plot or histogram, or a box plot.

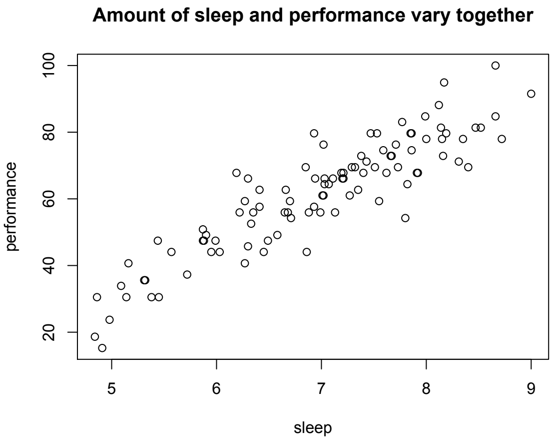
Variance



Covariance

When we have two variables, we can ask whether they are independent, or whether they *vary together* (potentially involving causality, where variance along one dimension affects variance along the other dimension).
A *scatterplot* can help to give us a visual impression of this.

Two variables are varying together

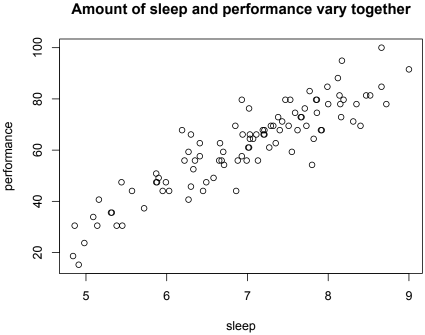


We call the joint variance "covariance".

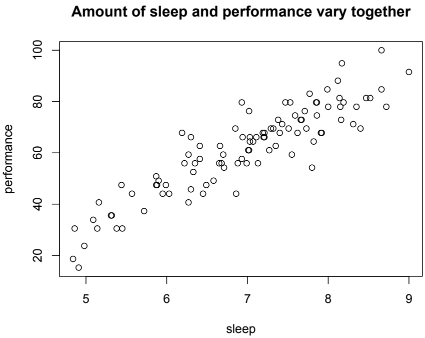
How can we quantify the *covariance*?

Notation: cov_{XY} or s_{XY}

$$cov_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N - 1}$$



How can we quantify the *covariance*?



Notation: cov_{XY} or s_{XY}

$$cov_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N - 1}$$

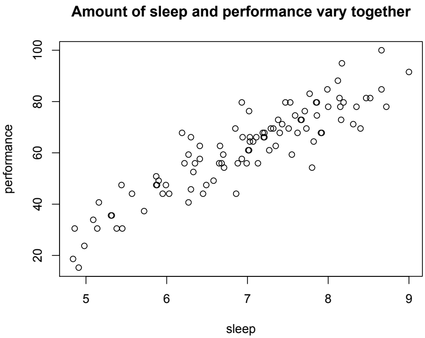
Does this do the right thing intuitively?

- if both X and Y are larger than their means, we'll get a high cov.

How can we quantify the *covariance*?

Notation: cov_{XY} or s_{XY}

$$cov_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N - 1}$$



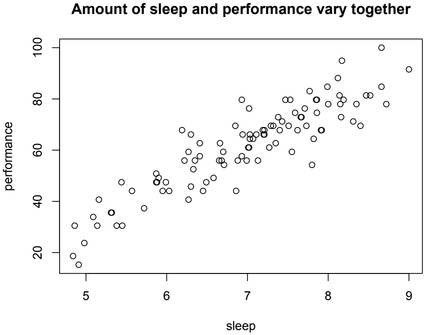
Does this do the right thing intuitively?

- if both X and Y are larger than their means, we'll get a high cov.
- if both X and Y are smaller than their means, we'll get a high cov.

How can we quantify the *covariance*?

Notation: cov_{XY} or s_{XY}

$$cov_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{N-1}$$



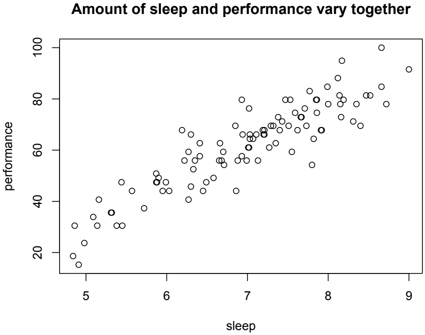
Does this do the right thing intuitively?

- if both X and Y are larger than their means, we'll get a high cov.
- if both X and Y are smaller than their means, we'll get a high cov.
- if X is larger than $\bar{X}$ and Y is sometimes larger and sometimes smaller than $\bar{Y}$ then we'll get a cov close to zero

How can we quantify the *covariance*?

Notation: cov_{XY} or s_{XY}

$$cov_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N - 1}$$



Does this do the right thing intuitively?

- if both X and Y are larger than their means, we'll get a high cov.
- if both X and Y are smaller than their means, we'll get a high cov.
- if X is larger than $\bar{X}$ and Y is sometimes larger and sometimes smaller than $\bar{Y}$ then we'll get a cov close to zero
- if X is larger than $\bar{X}$ when Y is lower than $\bar{Y}$, we'll get a negative cov.

Table of Contents

Pearson's product-moment correlation coefficient (r)

- Can we use *covariance* as a measure of the degree of relationship between two variables?

Pearson's product-moment correlation coefficient (r)

- Can we use *covariance* as a measure of the degree of relationship between two variables?
- **Difficulty:** the absolute value of is also a function of the standard deviations of X and Y.
- **Example:** a value of $cov_{XY} = 1.36$, for example, might reflect a high degree of correlation when the standard deviations are small, but a low degree of correlation when the standard deviations are high.

Pearson's product-moment correlation coefficient (r)

- Can we use *covariance* as a measure of the degree of relationship between two variables?
- **Difficulty:** the absolute value of is also a function of the standard deviations of X and Y.
- **Example:** a value of $cov_{XY} = 1.36$, for example, might reflect a high degree of correlation when the standard deviations are small, but a low degree of correlation when the standard deviations are high.
- **Pearson's product-moment correlation coefficient** r solves this by dividing covariance by the standard deviations of X and Y.

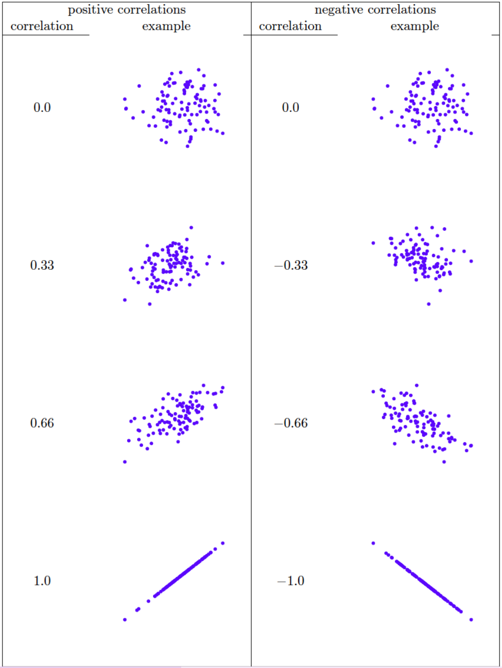
$$r = \frac{cov_{XY}}{s_X s_Y}$$

Pearson's correlation coefficient

$$r = \frac{cov_{XY}}{s_X s_Y}$$

- r ranges from -1 to 1.
- in R:

```
cor(sleep,perform)
[1] 0.903384
```



Correlations in R (with example from Navarro)

```
> summary(parenthood)
   dan.sleep       baby.sleep       dan.grump          day        
 Min.   :4.840   Min.   : 3.250   Min.   :41.00   Min.   :  1.00  
 1st Qu.:6.293   1st Qu.: 6.425   1st Qu.:57.00   1st Qu.: 25.75  
 Median :7.030   Median : 7.950   Median :62.00   Median : 50.50  
 Mean   :6.965   Mean   : 8.049   Mean   :63.71   Mean   : 50.50  
 3rd Qu.:7.740   3rd Qu.: 9.635   3rd Qu.:71.00   3rd Qu.: 75.25  
 Max.   :9.000   Max.   :12.070   Max.   :91.00   Max.   :100.00  
> cor(parenthood$dan.sleep, parenthood$baby.sleep)
[1] 0.6279493
> cor(parenthood)
             dan.sleep  baby.sleep   dan.grump         day
dan.sleep   1.00000000  0.62794934 -0.90338404 -0.09840768
baby.sleep  0.62794934  1.00000000 -0.56596373 -0.01043394
dan.grump  -0.90338404 -0.56596373  1.00000000  0.07647926
day        -0.09840768 -0.01043394  0.07647926  1.00000000
```

R and missing values

Imagine some data were missing from our data frame (i.e. on day one we did not record the duration of baby sleep):

	dan.sleep	baby.sleep	dan.grump	day
1	7.59	NA	56	1
2	7.91	11.66	60	2
3	5.14	7.92	82	3
4	7.71	9.61	55	4

```
> cor( parenthood2 )
           dan.sleep baby.sleep dan.grump day
dan.sleep          1         NA        NA  NA
baby.sleep        NA          1        NA  NA
dan.grump         NA         NA         1  NA
day               NA         NA        NA   1
```

R and missing values

Imagine some data were missing from our data frame (i.e. on day one we did not record the duration of baby sleep):

	dan.sleep	baby.sleep	dan.grump	day
1	7.59	NA	56	1
2	7.91	11.66	60	2
3	5.14	7.92	82	3
4	7.71	9.61	55	4

```
> cor( parenthood2 )
           dan.sleep baby.sleep dan.grump day
dan.sleep          1         NA        NA  NA
baby.sleep        NA          1        NA  NA
dan.grump         NA         NA         1  NA
day               NA         NA        NA   1
```

```
> cor(parenthood2, use = "complete.obs")
             dan.sleep baby.sleep   dan.grump         day
dan.sleep   1.00000000  0.6394985 -0.89951468  0.06132891
baby.sleep  0.63949845  1.0000000 -0.58656066  0.14555814
dan.grump  -0.89951468 -0.5865607  1.00000000 -0.06816586
day         0.06132891  0.1455581 -0.06816586  1.00000000
```

R and missing values

Imagine some data were missing from our data frame (i.e. on day one we did not record the duration of baby sleep):

```
  dan.sleep baby.sleep dan.grump day
1      7.59         NA        56   1
2      7.91      11.66        60   2
3      5.14       7.92        82   3
4      7.71       9.61        55   4
```

```
> cor( parenthood2 )
           dan.sleep baby.sleep dan.grump day
dan.sleep          1         NA        NA  NA
baby.sleep        NA          1        NA  NA
dan.grump         NA         NA         1  NA
day               NA         NA        NA   1
```

```
> cor(parenthood2, use = "complete.obs")
             dan.sleep baby.sleep   dan.grump         day
dan.sleep   1.00000000  0.6394985 -0.89951468  0.06132891
baby.sleep  0.63949845  1.0000000 -0.58656066  0.14555814
dan.grump  -0.89951468 -0.5865607  1.00000000 -0.06816586
day         0.06132891  0.1455581 -0.06816586  1.00000000
```

```
> cor(parenthood2, use = "pairwise.complete.obs")
            dan.sleep  baby.sleep   dan.grump          day
dan.sleep   1.00000000  0.61472303 -0.903442442 -0.076796665
baby.sleep  0.61472303  1.00000000 -0.567802669  0.058309485
dan.grump  -0.90344244 -0.56780267  1.000000000  0.005833399
day        -0.07679667  0.05830949  0.005833399  1.000000000
```

Talking about correlations

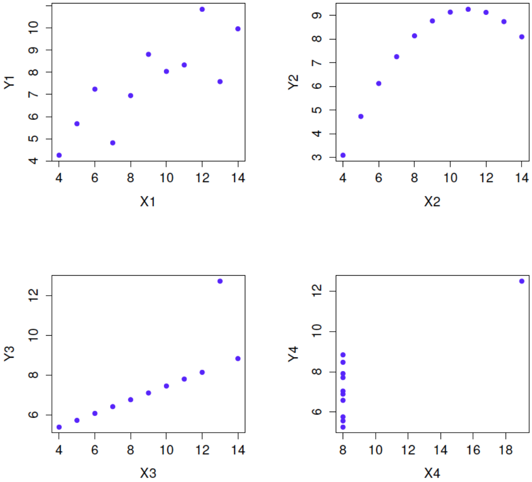
So how should you interpret a correlation of, say $r = .53$?

- a correlation coefficient **can not** be interpreted as a percentage: i.e. $r = 0.53$ doesn't mean that there's 53% of a relationship!
- Interpretation really depends on your type of data, your field etc.
  - there are cases where you should really expect correlations that strong. For instance, one of the benchmark data sets used to test theories of how people judge similarities is so clean that any theory that can't achieve a correlation of at least .9 really isn't deemed to be successful.
  - However, when looking for (say) elementary correlates of intelligence (e.g., inspection time, response time), if you get a correlation above .3 you're doing very very well.

Overview table

Correlation	Strength	Direction
-1.0 to -0.9	Very strong	Negative
-0.9 to -0.7	Strong	Negative
-0.7 to -0.4	Moderate	Negative
-0.4 to -0.2	Weak	Negative
-0.2 to 0	Negligible	Negative
0 to 0.2	Negligible	Positive
0.2 to 0.4	Weak	Positive
0.4 to 0.7	Moderate	Positive
0.7 to 0.9	Strong	Positive
0.9 to 1.0	Very strong	Positive

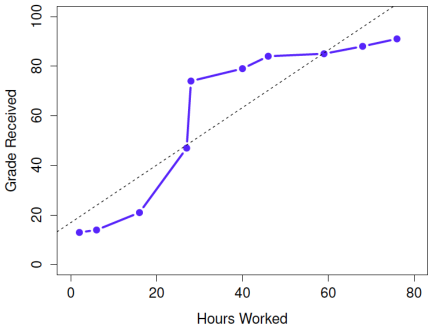
Interpreting Pearson correlation coefficient – what does your data actually look like? $r = .816$



What if the relation between the variables is not *linear*?

Example: Studying for an exam

Not a linear relationship. for a 10% improvement in grade you have to put in more time at the high end of the grading scale than at the low end of the grading scale.

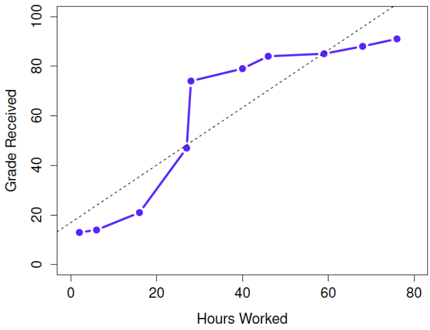


Pearson correlation: $r = 0.91$
(correlation would be 1 if all points fell on the dashed line.)

What if the relation between the variables is not *linear*?

Example: Studying for an exam

Not a linear relationship. for a 10% improvement in grade you have to put in more time at the high end of the grading scale than at the low end of the grading scale.



Pearson correlation: $r = 0.91$ (correlation would be 1 if all points fell on the dashed line.)

But we also see that more effort always leads to a better grade in this example.
This is what *Spearman's rank correlation* captures.

Spearman's rank correlation

instead of comparing the values of the two variables directly in the calculation of cov_{XY}, we convert all scores to **ranks**.

```
> effort
   hours grade
1      2    13
2     76    91
3     40    79
4      6    14
5     16    21
6     28    74
7     27    47
8     59    85
9     46    84
10    68    88
```

	rank (hours worked)	rank (grade received)
student 1	1	1
student 2	10	10
student 3	6	6
student 4	2	2
student 5	3	3
student 6	5	5
student 7	4	4
student 8	8	8
student 9	7	7
student 10	9	9

Spearman's rank correlation

instead of comparing the values of the two variables directly in the calculation of cov_{XY}, we convert all scores to **ranks**.

```
> effort
   hours grade
1      2    13
2     76    91
3     40    79
4      6    14
5     16    21
6     28    74
7     27    47
8     59    85
9     46    84
10    68    88
```

	rank (hours worked)	rank (grade received)
student 1	1	1
student 2	10	10
student 3	6	6
student 4	2	2
student 5	3	3
student 6	5	5
student 7	4	4
student 8	8	8
student 9	7	7
student 10	9	9

If two values are the same, they get the same rank by calculating the average rank for all identical values, e.g. if stud 9 would received a grade of 85, the rank in grades for both student 8 and 9 would have been 7.5

Spearman's rank correlation in R

```
hours.rank <- rank(effort$hours)
grade.rank <- rank(effort$grade)
```

Spearman's rank correlation in R

```
hours.rank <- rank(effort$hours)
grade.rank <- rank(effort$grade)

cor(hours,grade)
[1] 0.91
```

Spearman's rank correlation in R

```
hours.rank <- rank(effort$hours)
grade.rank <- rank(effort$grade)

cor(hours,grade)
[1] 0.91

cor(hours.rank,grade.rank)
[1] 1
```

Spearman's rank correlation in R

```
hours.rank <- rank(effort$hours)
grade.rank <- rank(effort$grade)

cor(hours,grade)
[1] 0.91

cor(hours.rank,grade.rank)
[1] 1

cor(hours,grade,method="spearman")
[1] 1
```

Kendall's τ as an alternative to Spearman's ρ

Kendall's τ is also appropriate for *ordinal* data (or when the exact distance in ranks is not important).

- for each observation, we have a pair of X and Y values
- sort these observations by the value of X and calculate the ranks for X and Y (as for Spearman's ρ)
- compare each pairs to all following pairs (N*(N-1)/2 comparisons), and check for which ones the ordering according to both X and Y is "concordant" (C: $x_i < x_j$ and $y_i < y_j$) vs. "discordant" (D)
- there can also be ties where the X values of two pairs are identical (T_X) or where the Y values in two pairs are identical (T_Y). Pairs that are entirely identical are not counted (T_{XY}).

$$\tau = \frac{C - D}{\sqrt{(C + D + T_X) \times (C + D + T_Y)}}$$

in R: `cor(hours, grade, method="kendall")`

Kendall's τ as an alternative to Spearman's ρ

$$\tau = \frac{C - D}{\sqrt{(C + D + T_X) \times (C + D + T_Y)}}$$

concordant C: $x_i < x_j$ and $y_i < y_j$
discordant D: $x_i < x_j$ and $y_i > y_j$
tie T_X :$x_i = x_j$ and $y_i \neq y_j$
tie T_Y :$x_i \neq x_j$ and $y_i = y_j$

Example data (height and weight):

cm	kg
1,68	60
1,68	63
1,72	58
1,82	82

Table of Contents

Significance tests for correlations

What we have done so far is to *describe* the relation between two ratio or interval scale variables. But there can be cases, where you also want to do *hypothesis testing* on the correlation.

So what are our hypotheses?

Significance tests for correlations

What we have done so far is to *describe* the relation between two ratio or interval scale variables. But there can be cases, where you also want to do *hypothesis testing* on the correlation.

So what are our hypotheses?

$H_0 : r = 0$
$H_1 : r \neq 0$

The following can be shown:
If the true correlation between two variables in the population is 0, and we sample with sufficiently large N, then the correlations found in the samples will be approximately normally distributed around 0.

Therefore, we can use the t-statistic.

one-tailed t-test

$$\frac{\bar{X}-\mu}{\frac{\hat{\sigma}}{\sqrt{N-1}}}$$

t-test for correlation

$$\frac{r}{\frac{\sqrt{1-r^2}}{\sqrt{N-2}}}$$

```
> cor.test(sleep,perform)

	Pearson's product-moment correlation

data:  sleep and perform
t = 20.854, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8594714 0.9340614
sample estimates:
     cor
0.903384
```

Are two correlations signif. different from one another?

When we test the difference between two independent rs, a minor difficulty arises. When the correlation in the population $\rho \neq 0$, the sampling distribution of r is not approximately normal (it becomes more and more skewed as $\rho \Rightarrow \pm 1$), and its standard error is not easily estimated. The same holds for the difference between correlations in two samples ($r_1 - r_2$).

Solution: Fisher's r-to-z transformation

Fisher (1921) showed that if we transform r to r', we can then validly use the t-test on r'.

$$r' = 0.5 \ \log_e \left| \frac{1+r}{1-r} \right|$$

(here: r' to avoid confusion with z)

(R has a function that does all the hard work for us.)

Significance tests for correlations: non-independent *r*s

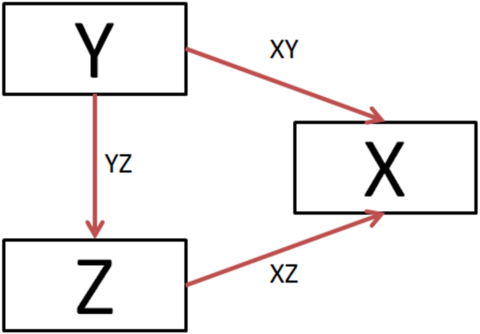
As usual, we have to distinguish between independent and non-independent (paired) observations.

Example

We have two computational models for predicting selectional preferences for a set of 240 verbs (i.e., whether *apple* is a good object for the verb *eat*).
Model A achieves a correlation of .65 with human judgments, model B achieves a correlation of .71 with the same set of human judgments. Both correlations are significantly different from zero, but is model B really significantly better than A?

Given that we estimated correlations for the same set of selectional preferences with both models, have a *paired correlation* in this case.

Paired correlations



A good overview is given here:
`http://www.philippsinger.info/?p=347`

paired.r {psych} R Documentation

Test the difference between (un)paired correlations

Description

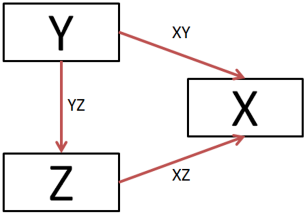
Test the difference between two (paired or unpaired) correlations. Given 3 variables, x, y, z, is the correlation between xy different than that between xz? If y and z are independent, this is a simple t-test of the z transformed rs. But, if they are dependent, it is a bit more complicated.

Usage

```
paired.r(xy, xz, yz=NULL, n, n2=NULL,twotailed=TRUE)
```

Arguments

`xy`	r(xy)
`xz`	r(xz)
`yz`	r(yz)
`n`	Number of subjects for first group
`n2`	Number of subjects in second group (if not equal to n)
`twotailed`	Calculate two or one tailed probability values



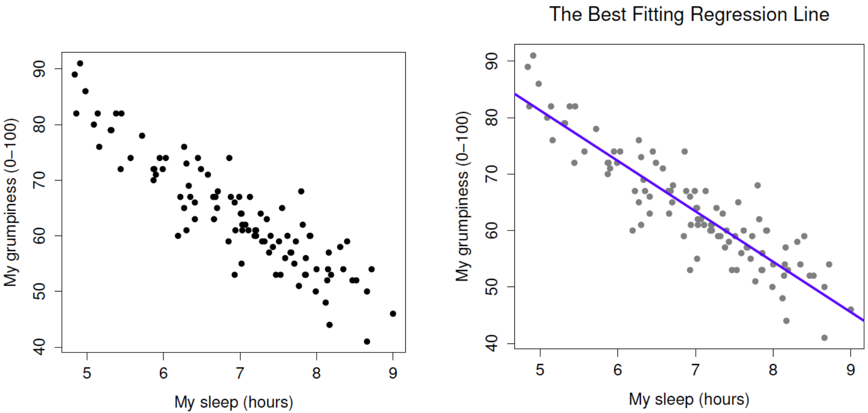

```
> paired.r(.65, .71, n=240)
Call: paired.r(xy = 0.65, xz = 0.71, n = 240)
[1] "test of difference between two independent correlations"
z = 1.22  With probability =  0.22
> paired.r(.65, .71, .8, n=240)
Call: paired.r(xy = 0.65, xz = 0.71, yz = 0.8, n = 240)
[1] "test of difference between two correlated  correlations"
t = -2.1  With probability =  0.04
> paired.r(.65, .71, .4, n=240)
Call: paired.r(xy = 0.65, xz = 0.71, yz = 0.4, n = 240)
[1] "test of difference between two correlated  correlations"
t = -1.34  With probability =  0.18
```

Table of Contents

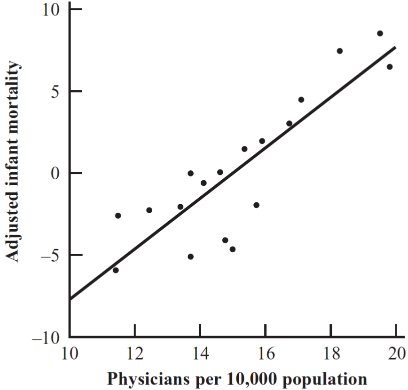
Linear Regression Models

We will spend most of the second half of the semester on linear (mixed effects) regression models. Pearson correlation is the simplest form of linear regression, but we will see that linear models are extremely powerful.

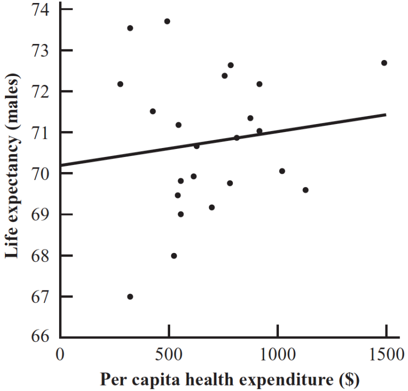
Regression line



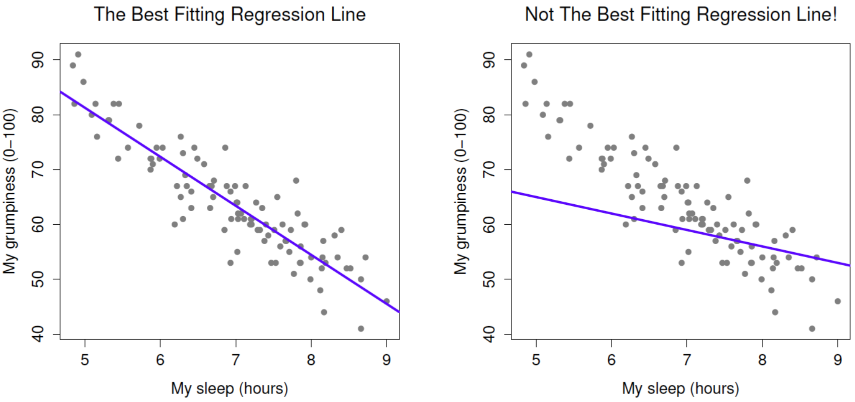
Other examples



(a) Infant mortality as a function of number of physicians



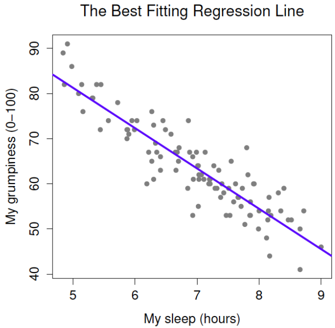
(b) Life expectancy as a function of health care expenditures



Regression line: $\hat{Y}_i = b_0 + b_1 X_i$

b_0 is the intercept
b_1 is the slope (how much increase / decrease in Y per unit of X)

A linear regression model

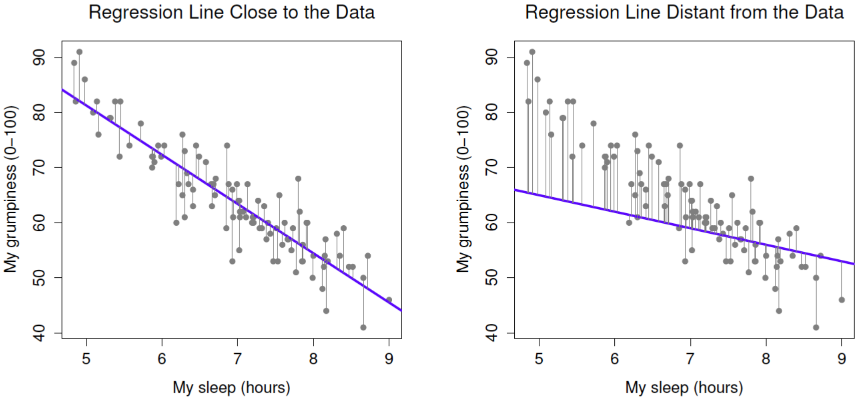


Not all points fall on the line – the difference between the prediction and the actual values of Y is the *error* or *residual*: $\epsilon_i = Y_i - \hat{Y}_i$

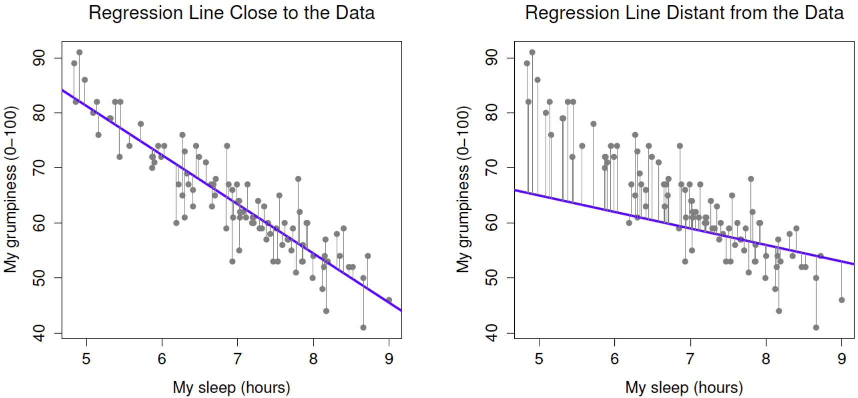
Regression model: $Y_i = b_0 + b_1 X_i + \epsilon_i$

Residuals

The residuals are a lot larger if the line doesn't fit the data well.



Ordinary Least Squares Regression



The estimated regression coefficients, $\hat{b}_0$ and $\hat{b}_1$ are those that minimise the sum of the squared residuals (= "ordinary least squares regression"). We can write this as $\sum_i (Y_i - \hat{Y}_i)^2$ or $\sum_i {\epsilon_i}^2$.

Estimating intercept and slope

Slope b_1:

$$b_1 = \frac{cov_{XY}}{s_X^2}$$

Intercept b_0:

$$b_0 = \bar{Y} - b_1\bar{X}$$

Estimating intercept and slope

Slope b_1:

$$b_1 = \frac{cov_{XY}}{s_X^2}$$

Intercept b_0:

$$b_0 = \bar{Y} - b_1\bar{X}$$

Compare the slope b_1 to the correlation coefficient r:

$$r = \frac{cov_{XY}}{s_X s_Y}$$

Correlation coefficient is "symmetric", while the regression line minimizes $\sum_i (Y_i - \hat{Y}_i)^2$. (If we instead regressed Y against X, the regression line would be different, as we would be minimizing $\sum_i (X_i - \hat{X}_i)^2$.)

Regressions in R

lm {stats} R Documentation

Fitting Linear Models

Description

`lm` is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although `aov` may provide a more convenient interface for these).

Usage

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

Arguments

`formula`
: an object of class "`formula`" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'.

`data`
: an optional data frame, list or environment (or object coercible by `as.data.frame` to a data frame) containing the variables in the model. If not found in `data`, the variables are taken from `environment(formula)`, typically the environment from which `lm` is called.

Regressions in R

We here need: `lm(formula, data)`
where `formula` is of the form: `response` $\sim$ `predictor`

Example

```
regression.1 <- lm(dan.grump ~ dan.sleep, data = parenthood)
```

```
> print( regression.1 )

Call:
lm(formula = dan.grump ~ dan.sleep, data = parenthood)

Coefficients:
(Intercept)    dan.sleep
    125.956       -8.937
```

Interpretation

```
> print( regression.1 )

Call:
lm(formula = dan.grump ~ dan.sleep, data = parenthood)

Coefficients:
(Intercept)    dan.sleep
    125.956       -8.937
```

Interpretation

Slope: for every hour of additional sleep, Dan will become less grumpy (reduce grumpiness by 8.937 on the grumpiness-scala). E.g., for 3h of sleep, Dan's grumpiness would be reduced by 26.811 compared to when he didn't get any sleep.

Intercept: level of grumpiness for 0 hours of sleep (125.956).

Table of Contents

Multiple Regression

Very often in science, we find that there is not just a single predictor but several predictors. Linear regression models can take into account several predictors.

formula for two predictors:

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \epsilon_i$$

general formula:

$$Y_i = b_0 + (\sum_{k=1}^{K} b_k X_{ik}) + \epsilon_i$$

Example in R

Example

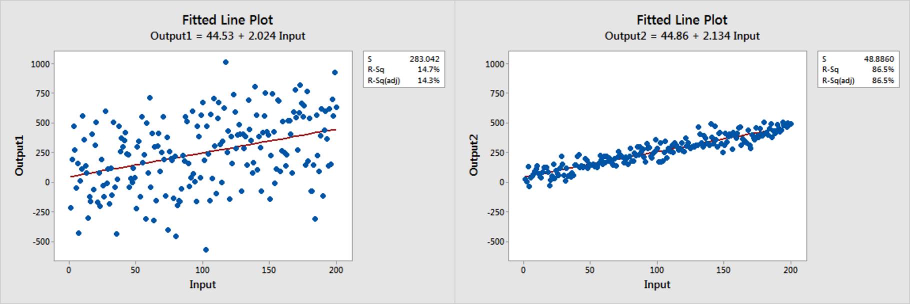
```
regression.2 <- lm(dan.grump ~ dan.sleep + baby.sleep,
    data = parenthood)
```

```
> print( regression.2 )

Call:
lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)

Coefficients:
(Intercept)    dan.sleep   baby.sleep
  125.96557     -8.95025      0.01052
```

We will spend a lot more time on more complicated regression models in later weeks.

Table of Contents

Quantifying the fit



Just knowing the regression line does not tell us how good the fit of the model is.

$$SS_{res} = \sum_i (Y_i - \hat{Y}_i)^2 \qquad SS_{tot} = \sum_i (Y_i - \bar{Y}_i)^2$$

We hope for small sum of squared residuals SS_{res} relative to our total variability SS_{tot}.

Calculating R^2 by hand

```
> # response: parenthood$dan.grump
> # predictor: parenthood$dan.sleep
> predicted_grump <- 125.97 + -8.94 * parenthood$dan.sleep
> SS.res <- sum( (parenthood$dan.grump - predicted_grump)^2 )
> SS.res
[1] 1838.722
> SS.tot <- sum( (parenthood$dan.grump - mean(parenthood
$dan.grump))^2 )
> SS.tot
[1] 9998.59
> Rsquared<- 1-(SS.res/SS.tot)
> Rsquared
[1] 0.8161018
> cor(parenthood$dan.grump , parenthood$dan.sleep )^2
[1] 0.8161027
```

Using the summary function in R:

```
> summary(lm(dan.grump~dan.sleep, data=parenthood))

Call:
lm(formula = dan.grump ~ dan.sleep, data = parenthood)

Residuals:
    Min      1Q  Median      3Q     Max
-11.025  -2.213  -0.399   2.681  11.750

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 125.9563     3.0161   41.76   <2e-16 ***
dan.sleep    -8.9368     0.4285  -20.85   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.332 on 98 degrees of freedom
Multiple R-squared:  0.8161, Adjusted R-squared:  0.8142
F-statistic: 434.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

Adjusted R-squared values

You'll often see *adjusted* R^2 reported: adjustment is an attempt to take the degrees of freedom into account.

$$\text{adj. } R^2 = 1 - \left(\frac{\mathrm{SS}_{res}}{\mathrm{SS}_{tot}} \times \frac{N-1}{N-K-1} \right)$$

(N= number of data points; K = number of predictors)

Adjusted R-squared values

You'll often see *adjusted* R^2 reported: adjustment is an attempt to take the degrees of freedom into account.

$$\text{adj. } R^2 = 1 - \left(\frac{\mathrm{SS}_{res}}{\mathrm{SS}_{tot}} \times \frac{N-1}{N-K-1} \right)$$

(N= number of data points; K = number of predictors)

advantage: when you add more predictors to the model, the adjusted R^2 value will only increase if the new variables improve the model performance more than you'd expect by chance.

disadvantage: the adjusted R^2 value can't be interpreted in the elegant way that R^2 can (proportion of variance that is explained by regression model).

Assumptions underlying linear regression

- Normality: residuals must be normally distributed.
- Linearity: relationship between X and Y should be linear!
- Homogeneity of variance: variance in Y doesn't change along scale of X (see next slide)
- Uncorrelated predictors
- Residuals are independent of each other
- no "bad" outliers (no small number of datapoints that have disproportionate influence on the model estimates)

(More on these questions in later weeks.)

Homogeneity of variance

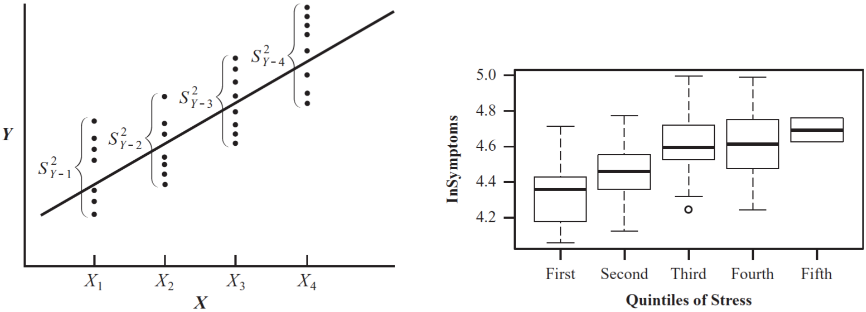


Figure 9.4 a) Scatter diagram illustrating regression assumptions; b) Similar plot for the data on Stress and Symptoms

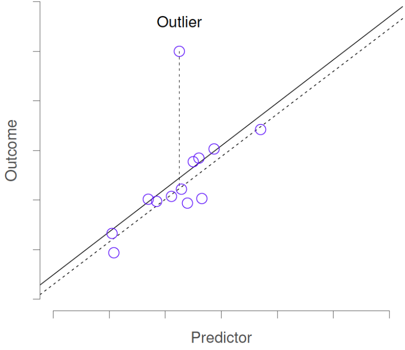
Spotting “bad outliers”

We want to identify outlier or other data points that are unusual or have disproportionate influence on the regression line.

We distinguish:

- **outlier**: observation that is very different from what the regression model predicts
- **high leverage**: an observation is unusual, but may be consistent with overall pattern.
- **influence of an observation**: outlier with high leverage
  $\rightarrow$ Cook's distance

Examples:



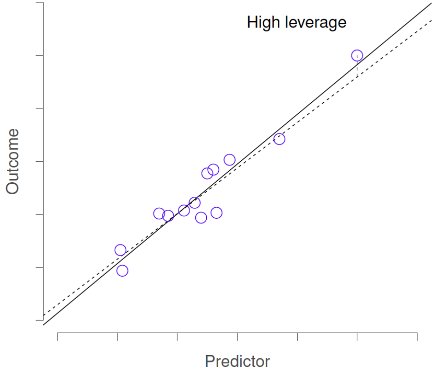
Spotting "bad outliers"

We want to identify outlier or other data points that are unusual or have disproportionate influence on the regression line.

We distinguish:

- **outlier**: observation that is very different from what the regression model predicts
- **high leverage**: an observation is unusual, but may be consistent with overall pattern.
- **influence of an observation**: outlier with high leverage
  → Cook's distance

Examples:



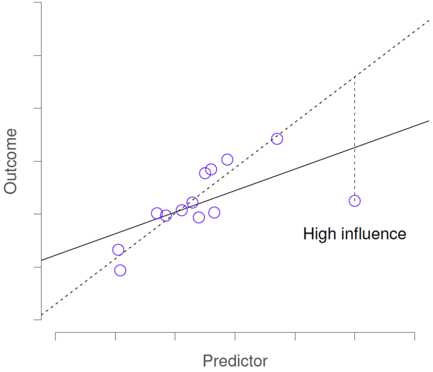
Spotting “bad outliers”

We want to identify outlier or other data points that are unusual or have disproportionate influence on the regression line.

We distinguish:

- **outlier**: observation that is very different from what the regression model predicts
- **high leverage**: an observation is unusual, but may be consistent with overall pattern.
- **influence of an observation**: outlier with high leverage
  $\rightarrow$ Cook's distance

Examples:



Outliers

How did the outlier come about?

Inspect your data / procedure for acquiring the data to see whether the outlier is a justified data point that should be in the data set or not.

Examples for reasons leading to outliers:

- extremely long reaction times due to fatigue or distraction
- extremely short reaction times due to inadvertently pushing a button
- words with zero frequency due to corpus limitations
- data input errors

Dealing with outliers

There are the following options:

1. removal
   - drop any value that is below or above an absolute cut-off
   - drop any value that is outside some range centered around the mean; usually
     - (m - 3sd,m + 3sd) or
     - (m - 2.5sd,m + 2.5sd)
2. substitution
   - replace any value that is outside the range [m - 3sd,m + 3sd] with the cut-off values (m - 3sd) and (m + 3sd) themselves.
3. Keep the data point because it is revealing about the data (but then check whether conclusions change when excluding the data point, and report accordingly)

Always report how you dealt with outliers, and what percentage of data points was removed overall.

Summary

- Correlation describes the relationship between two ratio or interval scale variables.
- Linearity of relationship or monotonicity? Pearson's r vs. Spearman's ρ.
- Hypothesis testing for correlation coefficients using t distribution.
- r-to-z transformation for comparing two correlations unequal to 0.
- Adjustment for paired correlations
- Correlation as a simple form of linear regression
- Linear regression describes one variable as a function of the other one.
- R^2 describes how much of the variance is explained by the linear regression model.