# Chapter 1 Overview of Statistical Data Science

## 1.1 What is data science?

Data science is an interdisciplinary field.

- It uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.
- It involves techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, domain knowledge and information science.
- It is related to data mining, machine learning and big data.

People have very different views regarding data science and statistics.

- Some argued that data science is not a new field, but rather another name for statistics.
- Some see that data science is applied statistics.
- Some see that data science as a brand new field that uses statistics.

Everyone agrees that statistics is a crucial component of data science.

Compared with *traditional* statistics (e.g., in the 70s), data science

- deals with new types of data (e.g., images, electronic health record),
- deals with huge datasets,
- and emphasizes prediction and action.

A crude schematic of a common data science project is show below. ds1 We can roughly categorize tasks of a data scientist based on the schematic.

| Task | Descriptions | Skills required |
| --- | --- | --- |
| Visioning | To generate hypotheses or questions that are of interest | Domain knowledge, self-learning, quantitative methods, etc. |
| Data acquisition | To gather data for verifying hypotheses via experiments, sample surveys, or data queries | Experimental designs, causal inference, query languages, etc. |
| Analysis | To analyze data to produce meaningful visualizations, build predictive models, test hypotheses, etc. | Domain knowledge, programming languages, statistical models, statistical theory, optimization, etc. |
| Conclusion | To conclude the analysis results in writing or via presentations | Communication skills, writing skills, public speaking skills, etc. |
| Management | To manage and oversee the project | Leadership, collaboration skills, tools for team projects, etc. |

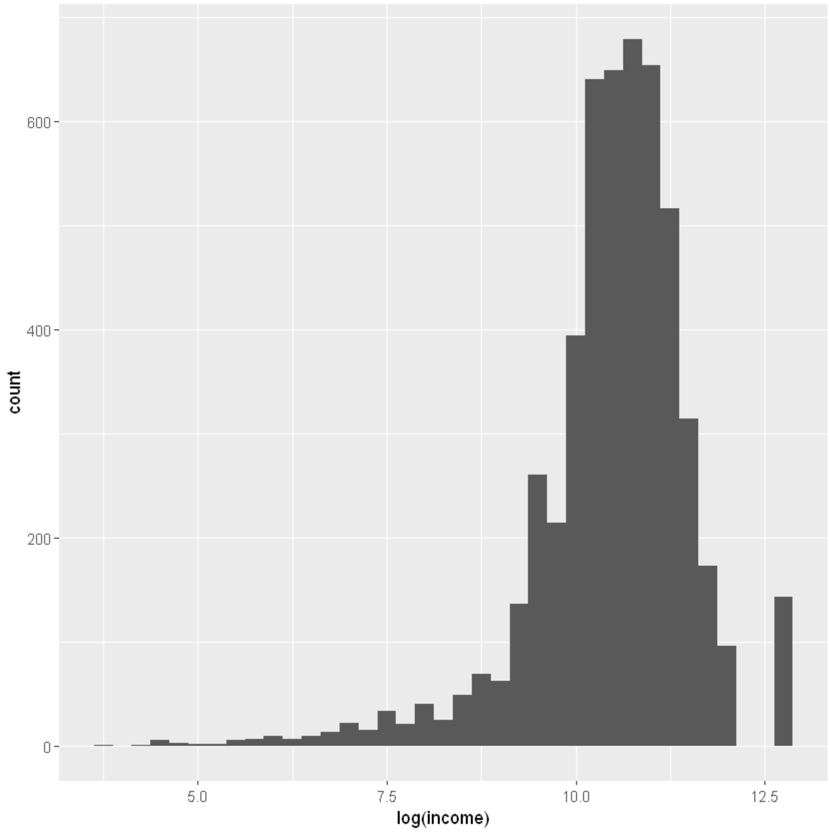
This collection of notes focuses on the **Data** part of the chain. ds2 The transform-visualise-model step is in fact an iterative process, where we need to reflect and revise our approaches based on results in previous steps. ds3

## 1.2 Example: `wages`

We now turn to the `wages` data to see an example of data analysis.

In [4]:
```
library(readxl)
library(ggplot2)
suppressPackageStartupMessages(library(tidyverse))

wages <- read_excel("../Data/wages.xlsx",na="NA")
wages %>%
  ggplot(aes(log(income))) + geom_histogram(binwidth = 0.25)
```




In [9]:
```
# Fit a linear regression using lm()
mod_e <- lm(log(income)~ education, data= wages)
```

```
Call:
lm(formula = log(income) ~ education, data = .)

Coefficients:
(Intercept)    education
     8.5577       0.1418
```

'lm'

In the code above, we propose a linear model for the response (log income) and the single predictor (education).

lm

The formula for `lm()` only needs to include the response (variable on the y axis) and predictors (variable on the x-axis). The intercept term is included by default, unless specified otherwise ( `-1` ).

lm

```
In [16]: mod_e
```

```
Call:
lm(formula = log(income) ~ education, data = .)

Coefficients:
(Intercept)    education
     8.5577       0.1418
```

**(Intercept):** 8.558

We can try to interpret the fitted coefficients.

- The average log income is {{as.numeric(round(mod_e$coef[1],digits=3))}} for those with zero years of education.

- The average difference in log income is {{as.numeric(round(mod_e$coef[2],digits=3))}} for groups with one year difference in education.

Neither statement makes a lot of sense. For instance, (1) there isn't anyone with zero years of education in the `wages` dataset, and (2) differences in log income are not informative for general audience. Therefore, we have two questions to think about.

- Why take the logarithm of income?
- How should we interpret the fitted results?

After model fitting, we usually proceed to hypothesis testing, predictive modeling, or model diagnostics. An experienced reader can certainly perform these tasks from scratch. However, it is best not to reinvent the wheels. For very common tasks, it is extremely likely that there are existing tools out there on the Internet. Here we use the package `broom` in `R`.

`Broom` includes three functions which work for most types of models (and can be extended to more):

1. `tidy()` - returns model coefficients, stats: what uncertainty is associated with it?
2. `glance()` - returns model diagnostics: how "good" is the model?
3. `augment()` - returns predictions, residuals, and other raw values

```
In [30]: library(broom)
         mod_e %>% tidy()
```

A tibble: 2 × 5

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> |
| (Intercept) | 8.5576906 | 0.073259622 | 116.81320 | 0.000000e+00 |
| education | 0.1418404 | 0.005304577 | 26.73924 | 8.408952e-148 |

```
In [31]: mod_e %>% glance()
```

A tibble: 1 × 12

| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | devia |
|---|---|---|---|---|---|---|---|---|---|
| <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <d |

| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | devia |
|---|---|---|---|---|---|---|---|---|---|
| <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <d |
| 0.1196233 | 0.119456 | 0.9923358 | 714.987 | 8.408952e-148 | 1 | -7427.793 | 14861.59 | 14881.29 | 5181. |

In [32]:

```
mod_e %>% augment()
```

A tibble: 5264 × 8

| .rownames | log(income) | education | .fitted | .std.resid | .hat | .sigma | .cooksd |
|---|---|---|---|---|---|---|---|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | 9.852194 | 13 | 10.401615 | -0.55371967 | 0.0001991827 | 0.9924012 | 3.054133e-05 |
| 2 | 10.463103 | 10 | 9.976094 | 0.49090632 | 0.0005537086 | 0.9924074 | 6.675581e-05 |
| 3 | 11.561716 | 16 | 10.827137 | 0.74038545 | 0.0003590043 | 0.9923784 | 9.843315e-05 |
| 4 | 10.596635 | 14 | 10.543456 | 0.05359492 | 0.0001953068 | 0.9924299 | 2.805560e-07 |
| 5 | 11.225243 | 14 | 10.543456 | 0.68712042 | 0.0001953068 | 0.9923856 | 4.611455e-05 |
| 6 | 11.532728 | 18 | 11.110817 | 0.42532922 | 0.0007513008 | 0.9924131 | 6.800811e-05 |
| 7 | 11.156251 | 12 | 10.259775 | 0.90351685 | 0.0002602083 | 0.9923532 | 1.062372e-04 |
| 8 | 11.002100 | 12 | 10.259775 | 0.74815540 | 0.0002602083 | 0.9923774 | 7.284298e-05 |
| 9 | 11.918391 | 13 | 10.401615 | 1.52864202 | 0.0001991827 | 0.9922098 | 2.327661e-04 |
| 10 | 11.652687 | 16 | 10.827137 | 0.83207630 | 0.0003590043 | 0.9923648 | 1.243231e-04 |
| 11 | 12.747903 | 16 | 10.827137 | 1.93594843 | 0.0003590043 | 0.9920766 | 6.729971e-04 |
| 12 | 10.596635 | 16 | 10.827137 | -0.23232373 | 0.0003590043 | 0.9924251 | 9.691986e-06 |
| 13 | 10.463103 | 14 | 10.543456 | -0.08098093 | 0.0001953068 | 0.9924295 | 6.405274e-07 |
| 14 | 11.561716 | 16 | 10.827137 | 0.74038545 | 0.0003590043 | 0.9923784 | 9.843315e-05 |
| 15 | 9.682591 | 16 | 10.827137 | -1.15359186 | 0.0003590043 | 0.9923046 | 2.389626e-04 |
| 16 | 8.556414 | 12 | 10.259775 | -1.71674017 | 0.0002602083 | 0.9921522 | 3.835423e-04 |
| 17 | 10.645425 | 12 | 10.259775 | 0.38867895 | 0.0002602083 | 0.9924159 | 1.966012e-05 |

| .rownames | log(income) | education | .fitted | .std.resid | .hat | .sigma | .cooksd |
|---:|---:|---:|---:|---:|---:|---:|---:|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 18 | 10.878047 | 12 | 10.259775 | 0.62312838 | 0.0002602083 | 0.9923935 | 5.053115e-05 |
| 19 | 10.558414 | 12 | 10.259775 | 0.30098414 | 0.0002602083 | 0.9924216 | 1.178939e-05 |
| 20 | 10.691945 | 12 | 10.259775 | 0.43556436 | 0.0002602083 | 0.9924122 | 2.468930e-05 |
| 21 | 11.440355 | 12 | 10.259775 | 1.18985260 | 0.0002602083 | 0.9922966 | 1.842428e-04 |
| 22 | 11.034890 | 12 | 10.259775 | 0.78120277 | 0.0002602083 | 0.9923726 | 7.942032e-05 |
| 23 | 10.714418 | 14 | 10.543456 | 0.17229923 | 0.0001953068 | 0.9924273 | 2.899606e-06 |
| 24 | 11.775290 | 16 | 10.827137 | 0.95564771 | 0.0003590043 | 0.9923440 | 1.639915e-04 |
| 25 | 10.714418 | 12 | 10.259775 | 0.45821373 | 0.0002602083 | 0.9924103 | 2.732375e-05 |
| 26 | 11.225243 | 12 | 10.259775 | 0.97305163 | 0.0002602083 | 0.9923408 | 1.232185e-04 |
| 27 | 11.198215 | 14 | 10.543456 | 0.65988033 | 0.0001953068 | 0.9923891 | 4.253071e-05 |
| 28 | 10.915088 | 12 | 10.259775 | 0.66046059 | 0.0002602083 | 0.9923890 | 5.676726e-05 |
| 29 | 12.747903 | 14 | 10.543456 | 2.22168956 | 0.0001953068 | 0.9919646 | 4.821021e-04 |
| 30 | 10.308953 | 12 | 10.259775 | 0.04956390 | 0.0002602083 | 0.9924299 | 3.196943e-07 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 5237 | 10.714418 | 12 | 10.259775 | 0.4582137 | 0.0002602083 | 0.9924103 | 2.732375e-05 |
| 5238 | 9.392662 | 12 | 10.259775 | -0.8739238 | 0.0002602083 | 0.9923581 | 9.939196e-05 |
| 5239 | 7.600902 | 12 | 10.259775 | -2.6797567 | 0.0002602083 | 0.9917527 | 9.345334e-04 |
| 5240 | 9.472705 | 14 | 10.543456 | -1.0791263 | 0.0001953068 | 0.9923203 | 1.137409e-04 |
| 5241 | 10.043249 | 12 | 10.259775 | -0.2182262 | 0.0002602083 | 0.9924256 | 6.197521e-06 |
| 5242 | 9.190138 | 18 | 11.110817 | -1.9362412 | 0.0007513008 | 0.9920765 | 1.409383e-03 |
| 5243 | 10.545341 | 12 | 10.259775 | 0.2878094 | 0.0002602083 | 0.9924223 | 1.077988e-05 |
| 5244 | 9.615805 | 11 | 10.117935 | -0.5061031 | 0.0003783836 | 0.9924060 | 4.847799e-05 |

| .rownames | log(income) | education | .fitted | .std.resid | .hat | .sigma | .cooksd |
|---|---|---|---|---|---|---|---|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 5245 | 10.950807 | 18 | 11.110817 | -0.1613071 | 0.0007513008 | 0.9924277 | 9.781768e-06 |
| 5246 | 11.982929 | 16 | 10.827137 | 1.1649283 | 0.0003590043 | 0.9923022 | 2.436823e-04 |
| 5247 | 12.747903 | 18 | 11.110817 | 1.6503493 | 0.0007513008 | 0.9921733 | 1.023910e-03 |
| 5248 | 10.705489 | 12 | 10.259775 | 0.4492150 | 0.0002602083 | 0.9924111 | 2.626108e-05 |
| 5249 | 11.082143 | 16 | 10.827137 | 0.2570217 | 0.0003590043 | 0.9924239 | 1.186220e-05 |
| 5250 | 10.165852 | 6 | 9.408733 | 0.7636642 | 0.0018265061 | 0.9923751 | 5.335683e-04 |
| 5251 | 9.118225 | 11 | 10.117935 | -1.0076213 | 0.0003783836 | 0.9923344 | 1.921593e-04 |
| 5252 | 10.373491 | 10 | 9.976094 | 0.4005770 | 0.0005537086 | 0.9924150 | 4.444920e-05 |
| 5253 | 10.434116 | 12 | 10.259775 | 0.1757101 | 0.0002602083 | 0.9924272 | 4.017887e-06 |
| 5254 | 9.210340 | 12 | 10.259775 | -1.0576774 | 0.0002602083 | 0.9923246 | 1.455830e-04 |
| 5255 | 9.998798 | 11 | 10.117935 | -0.1200798 | 0.0003783836 | 0.9924288 | 2.729018e-06 |
| 5256 | 10.596635 | 16 | 10.827137 | -0.2323237 | 0.0003590043 | 0.9924251 | 9.691986e-06 |
| 5257 | 10.933107 | 12 | 10.259775 | 0.6786206 | 0.0002602083 | 0.9923867 | 5.993192e-05 |
| 5258 | 9.952278 | 12 | 10.259775 | -0.3099125 | 0.0002602083 | 0.9924211 | 1.249921e-05 |
| 5259 | 9.998798 | 14 | 10.543456 | -0.5489182 | 0.0001953068 | 0.9924017 | 2.942982e-05 |
| 5260 | 9.546813 | 12 | 10.259775 | -0.7185624 | 0.0002602083 | 0.9923814 | 6.719439e-05 |
| 5261 | 8.006368 | 12 | 10.259775 | -2.2711068 | 0.0002602083 | 0.9919436 | 6.712422e-04 |
| 5262 | 11.184421 | 14 | 10.543456 | 0.6459791 | 0.0001953068 | 0.9923908 | 4.075766e-05 |
| 5263 | 10.043249 | 16 | 10.827137 | -0.7900831 | 0.0003590043 | 0.9923713 | 1.120911e-04 |
| 5264 | 11.350407 | 12 | 10.259775 | 1.0991979 | 0.0002602083 | 0.9923162 | 1.572374e-04 |
| 5265 | 9.798127 | 12 | 10.259775 | -0.4652740 | 0.0002602083 | 0.9924097 | 2.817226e-05 |
| 5266 | 10.126631 | 12 | 10.259775 | -0.1341897 | 0.0002602083 | 0.9924284 | 2.343379e-06 |

In [33]:

```
mod_e %>%
  tidy() %>% filter(p.value < 0.05)
```

A tibble: 2 × 5

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> |
| (Intercept) | 8.5576906 | 0.073259622 | 116.81320 | 0.000000e+00 |
| education | 0.1418404 | 0.005304577 | 26.73924 | 8.408952e-148 |

The low $R^2$ indicates that education explains only a part of variability of income. We can include more predictors in the model.

In [34]:

```
mod_eh <- wages %>%
  lm(log(income) ~ education + height, data = .)
mod_eh %>% tidy()
```

A tibble: 3 × 5

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> |
| (Intercept) | 5.34837618 | 0.231320415 | 23.12107 | 1.002503e-112 |
| education | 0.13871285 | 0.005205245 | 26.64867 | 7.120134e-147 |
| height | 0.04830864 | 0.003309870 | 14.59533 | 2.504935e-47 |

In [36]:

```
mod_eh %>% glance()
```

A tibble: 1 × 12

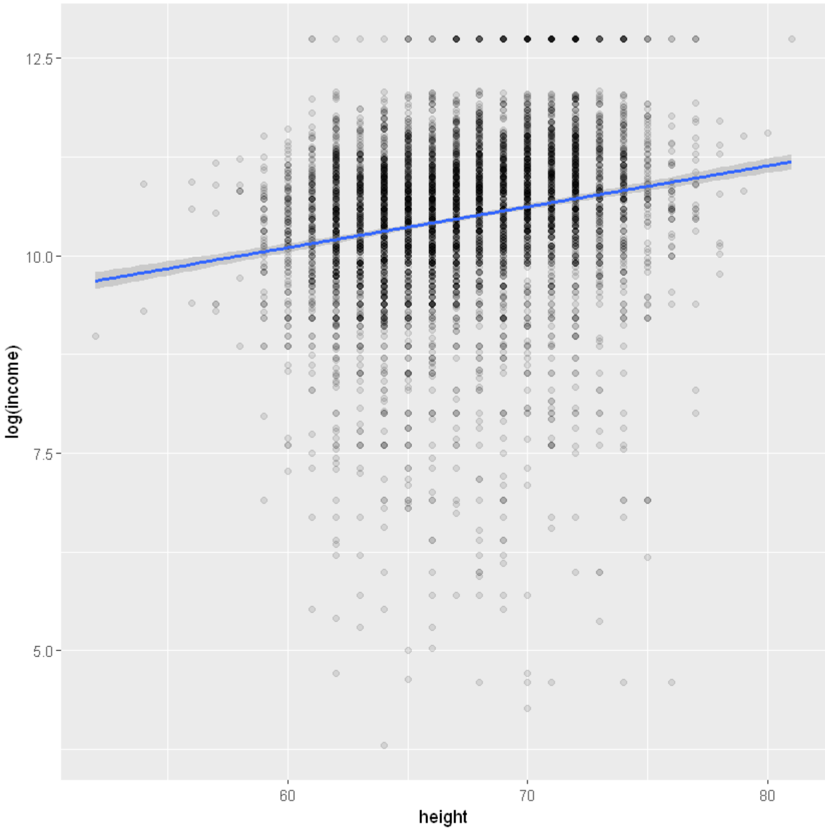
| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | devi |
|---|---|---|---|---|---|---|---|---|---|
| <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | < |
| 0.1538835 | 0.1535618 | 0.9729281 | 478.4099 | 1.273687e-191 | 2 | -7323.321 | 14654.64 | 14680.92 | 498 |

The $R^2$ does improves a bit, but still remains low. Maybe the linear model is a not a good choice. It might be a good idea to look at the raw data.

In [38]:

```
wages %>%
  ggplot(aes(x = height, y = log(income))) +
    geom_point(alpha = 0.1) +
    geom_smooth(method = lm)
```

```
`geom_smooth()` using formula 'y ~ x'
```




In [39]:
```
wages %>%
  ggplot(aes(x = height, y = log(income))) +
    geom_point(alpha = 0.1) +
    geom_smooth(method = loess)
```

```
`geom_smooth()` using formula 'y ~ x'
```