DAT 201 Final Assignment

Linear Regression of Life Expectancy and Term Life Insurance

Colin Bowers (colin.bowers@gmail.com)
April 15, 2023

Agenda

Two sets of data

1. Life Expectancy by Country (UN)
2. Term Life Insurance

For each dataset

1. Data Overview
2. Data Preparation
3. Modelling
4. Assess Performance
5. Summary of Results

Assumptions

- This presentation was prepared for a general audience with limited expertise in statistics
- More details of the analysis, along with a technical commentary, can be found in the accompanying R code

Part I

UN Life Expectancy

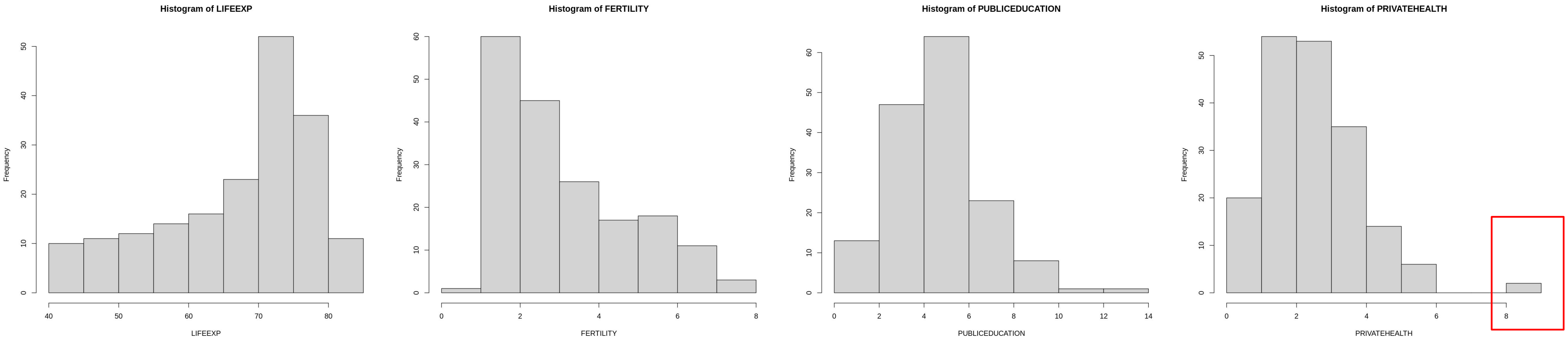
- Health care is a complex issue that affects everyone.
- Comparing different health care systems allows us to design better health care systems.
- We evaluated the correlation between life expectancy with average family size, public education spending and private health care spending.

Data Overview

File Name: UNLifeExpectancy	Number of obs: 185	Number of variables: 15
Variable	Number of Obs Missing	**Description**
REGION		Categorical variable for region of the world
COUNTRY		The name of the country
LIFEEXP		Life expectancy at birth, in years
ILLITERATE	14	Adult illiteracy rate, % aged 15 and older
POP	1	2005 population, in millions
FERTILITY	4	Total fertility rate, births per woman
PRIVATEHEALTH	1	2004 Private expenditure on health, % of GDP
PUBLICEDUCATION	28	Public expenditure on education, % of GDP
HEALTHEXPEND	5	2004 Health expenditure per capita, PPP in USD
BIRTHATTEND	7	Births attended by skilled health personnel (%)
PHYSICIAN	3	Physicians per 100,000 people
SMOKING	88	Prevalence of smoking, (male) % of adults
RESEARCHERS	95	Researchers in R & D, per million people
GDP	7	Gross domestic product, in billions of USD
FEMALEBOSS	87	Legislators, senior officials and managers, % female

Source: United Nations Human Development Report, available at http://hdr.undp.org/en/.

Data Overview



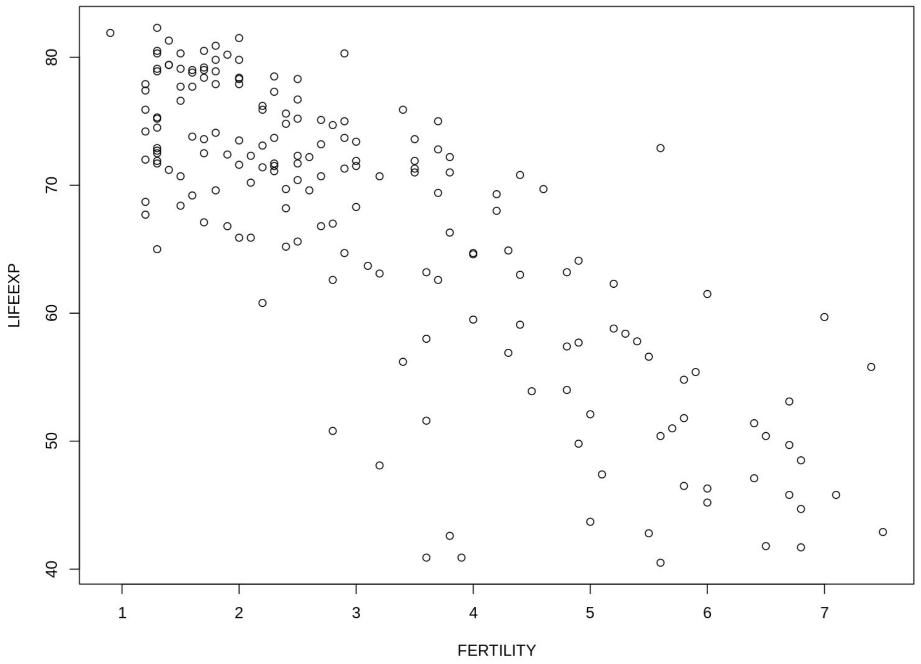
LIFEEXP	FERTILITY	PUBLICEDUCATION	PRIVATEHEALTH
Min. :40.50	Min. :0.90	Min. : 0.600	Min. :0.300
1st Qu.:59.70	1st Qu.:1.80	1st Qu.: 3.400	1st Qu.:1.500
Median :71.00	Median :2.70	Median : 4.600	Median :2.400
Mean :67.05	Mean :3.19	Mean : 4.695	Mean :2.517
3rd Qu.:75.10	3rd Qu.:4.30	3rd Qu.: 5.800	3rd Qu.:3.300
Max. :82.30	Max. :7.50	Max. :13.400	Max. :8.500
	NA's :4	NA's :28	NA's :1

Life Expectancy and Fertility

Visual inspection:

- Downward trend
- Implies that an increase in fertility leads to reduced life expectancy

Correlation (Pearson): **-0.806**



Life Expectancy and Fertility

Results of Linear Regression

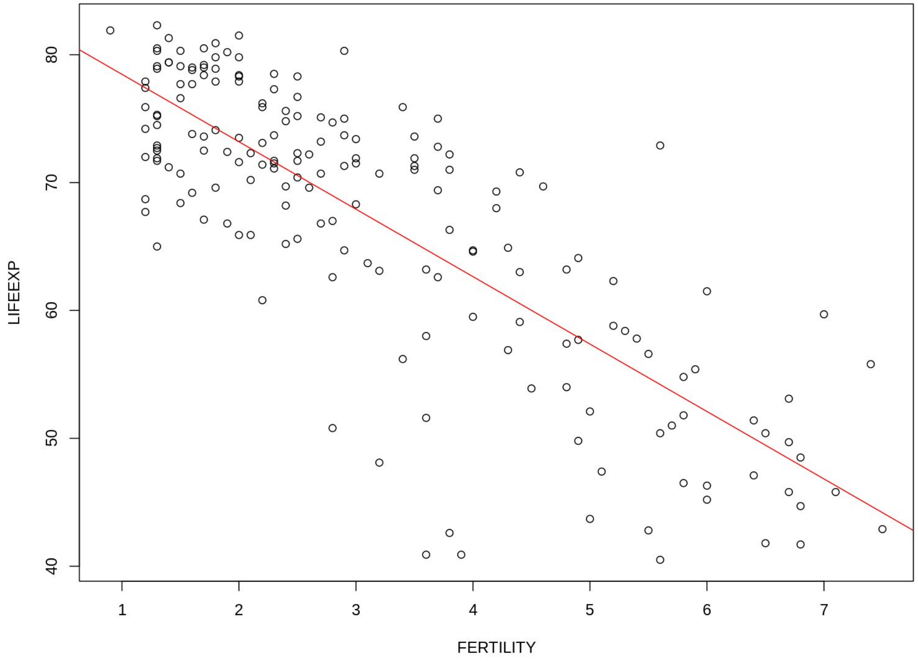
The fitted regression model was:

`LIFEEXP = 83.74 + (-5.27)FERTILITY`

The overall regression was statistically significant ($R2 = 0.645$, $F(1, 149) = 333.7$, $p < 2.2e-16$).

It was found that number of children significantly predicted life expectancy. ($\beta = -5.3993$, $p < 2e-16$).

The model predicts that for every additional child in a family, the life expectancy **drops by 5.4 years**.



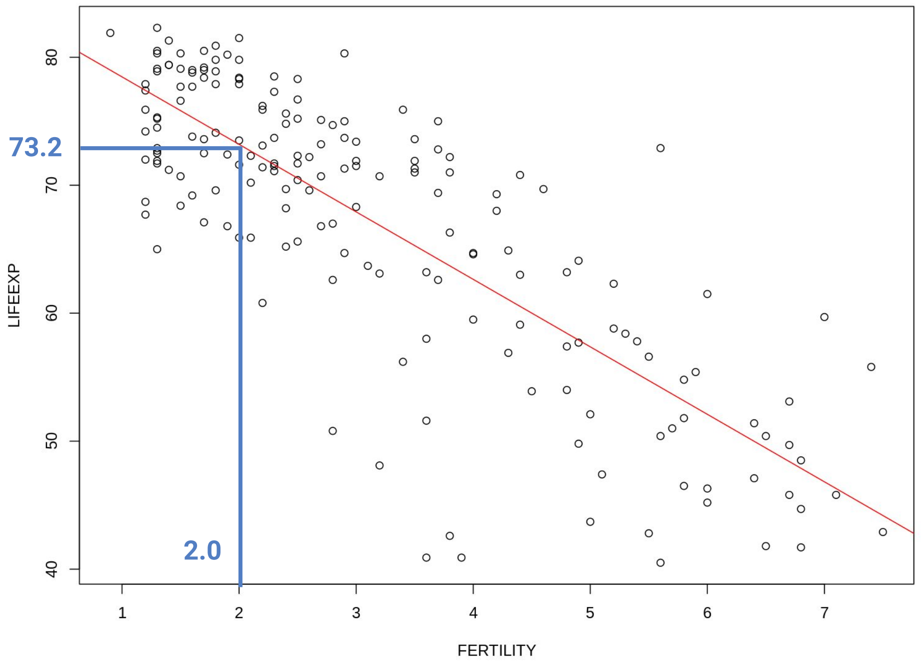
Life Expectancy and Fertility

Using the Model

For example, the **United States** has an average fertility rate of 2.0.

The fitted life expectancy for the United States is **73.2 (±1.2) years**.

The interval stated above is a 95% confidence interval which, given as a range, is 72.0 - 74.4 years.

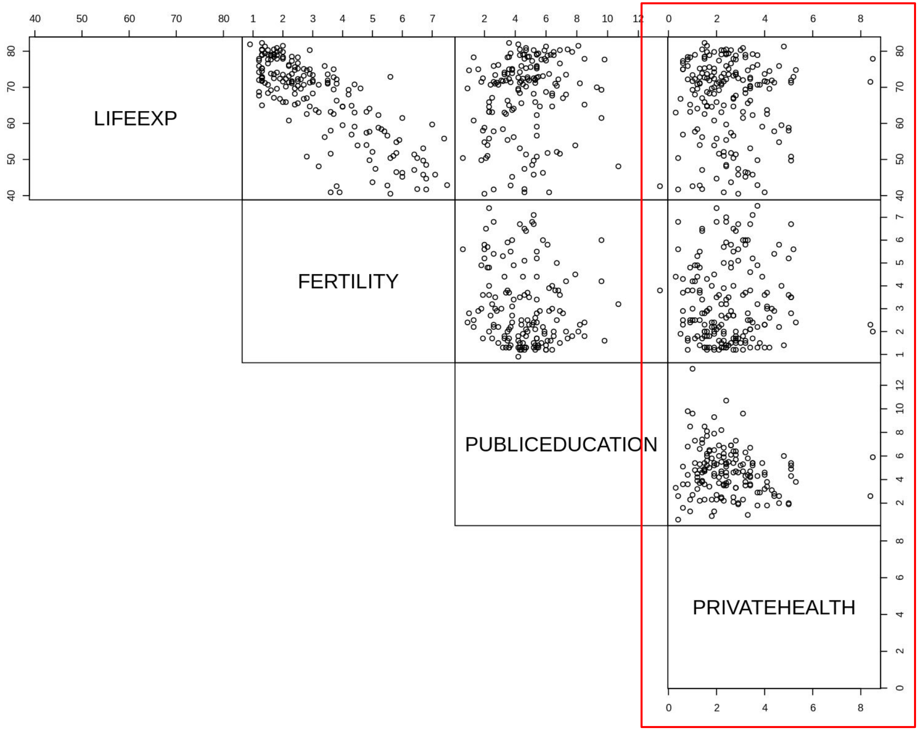


Fertility, Education & Private Healthcare

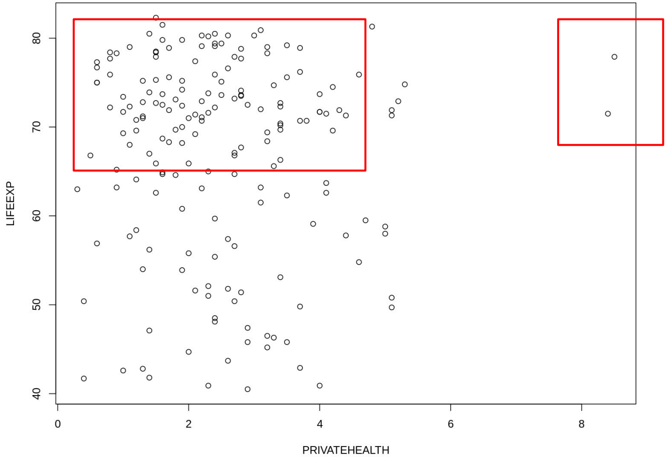
Private Health

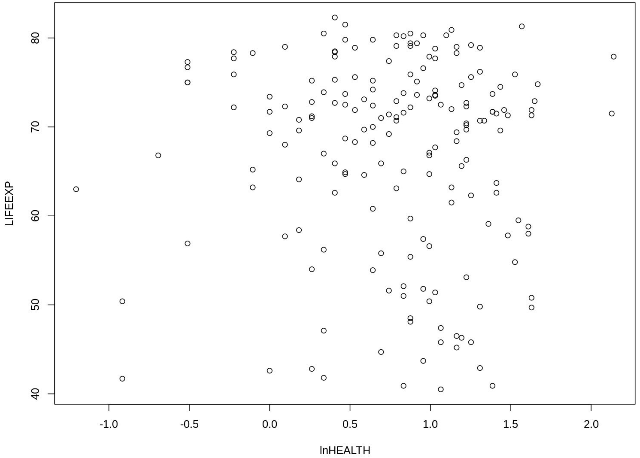
- Values for Private Health (highlighted red) are skewed
- That is, the plots show comes clustering towards the left.
- Difficult to determine a linear relationship

Let's take a closer look…



Transforming Private Healthcare





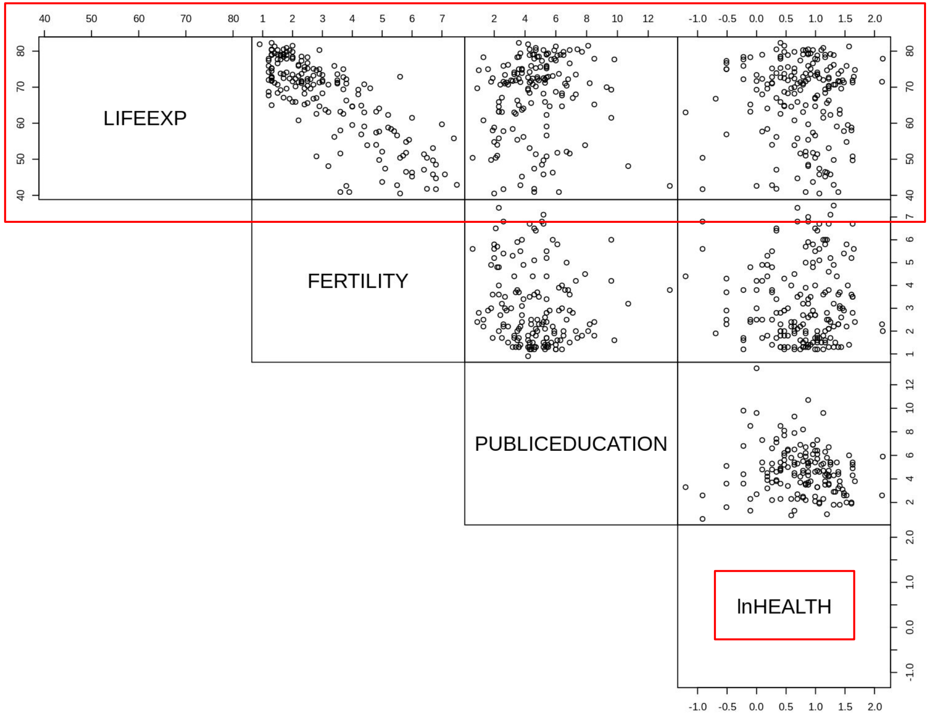
- There are a few outliers (top-right)
- Remaining data points are closely clustered together.
- This data is slightly skewed to the right as per the histogram and median < mean (from slide 5).

<u>Solution</u>: take the natural log transform, such that:

`lnHEALTH = log(PRIVATEHEALTH)`

Fertility, Education & Private Healthcare*

- Replaced PRIVATE HEALTH with the new lnHEALTH variable
- Top row shows the three variables we will use to create our model



Let's proceed with a multiple linear regression…

Data Modelling

Multiple linear regression

$$y = 85.62 + (-5.40)x_1 + (-0.18)x_2 + (-1.03)x_3$$

Interpretation

For every **additional child** in a family, the predicted life expectancy **drops by 5.4 years**.

For each additional percentage of GDP spent on **public education**, the predicted life expectancy **drops by 0.2 years**.

For a **1% change in health expenditure** the predicted change in life expectancy is **reduced by 0.01** years.

	Variable	Coefficient	Std. Err.
	(Intercept)	85.6264	2.0033
X1	FERTILITY	-5.3993	0.3308
X2	PUBLIC EDUCATION	-0.1846	0.2685
X3	lnHEALTH	-1.0296	0.9431

Assessing Performance

PERFORMANCE STATS

Multiple R-squared: 0.645

Adjusted R-squared: 0.6378

Residual standard error (RSE): 6.645

F-statistic: 89.64 on 3 and 148 DF

p-value: < 2.2e-16

CONCLUSIONS

The model is able to explain **64.5%** of the variance in life expectancy by using the three explanatory variables.

Any predicted value using this model will have a 95% chance to be **+/- 13.3 years** of the line (*2 * RSE*).

Assessing Significance

Variable	T-Value	P-Value	Significant?
(Intercept)	42.742	0.000	Yes
FERTILITY	-16.324	0.000	Yes
PUBLIC EDUCATION	-0.688	0.493	No
lnHEALTH	-1.092	0.277	No

Assuming a confidence threshold of 95%, any p-value greater than 0.05 will be considered NOT statistically significant.

CONCLUSIONS

It was found that **fertility** (p=0.000) significantly predicted values for **face**.

However, **public education** (p=0.493) and spending on **private health** (p=0.277) was found to NOT be a significant predictor.

Summary of Results

Multiple linear regression was used to test if **number of children, years of education** and **spending on health care** significantly predicted **life expectancy**.

The fitted regression model was:

```
LIFEEXP = 85.62 + (-5.40)FERTILITY + (-0.18)PUBLICEDUCATION + (-1.03)lnHEALTH
```

The overall regression was statistically significant ($R2 = 0.645$, $F(3, 148) = 89.64$, $p < 2e\text{-}16$).

It was found that **number of children** significantly predicted life expectancy. ($\beta = -5.3993$, $p < 2e\text{-}16$).
The model predicts that for every additional child in a family, the life expectancy **drops by about 5 years**.

It was found that the amount spent on **public education** did <u>not</u> significantly predict life expectancy ($\beta = -0.1846$, $p = 0.493$).

It was found that **amount spent on healthcare** did <u>not</u> significantly predict life expectancy ($\beta = -1.0296$, $p = 0.277$).
For a 1% change in health expenditure, if all other variables remain fixed, the predicted change in life expectancy is **reduced by 0.01 years**.

Part II

Term Life Insurance

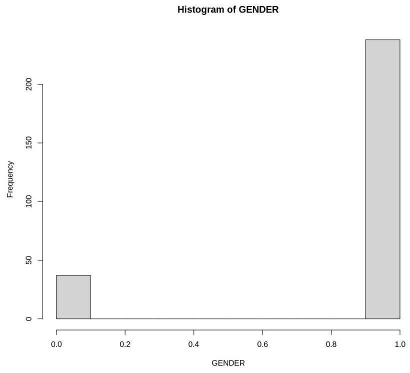
- Life insurance companies continually seek new ways to deliver products to the market.
- They wish to know “who buys insurance and how much do they buy?”
- We evaluated the correlation between insurance payout amount with gender, age and income of individuals that purchased insurance.

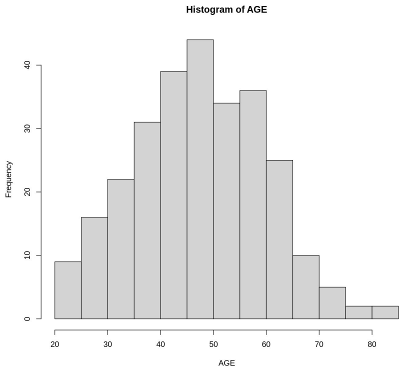
Data Overview

File Name: TermLife	Number of obs: 500	Number of variables: 18
Variable	Number of Obs Missing	**Description**
GENDER		Gender of the survey respondent
AGE		Age of the survey respondent
MARSTAT		Marital status of the survey respondent (=1 if married, =2 if living with partner, and =0 otherwise)
EDUCATION		Number of years of education of the survey respondent
ETHNICITY		Ethnicity
SMARSTAT		Marital status of the respondent's spouse
SGENDER		Gender of the respondent's spouse
SAGE		Age of the respondent's spouse
SEDUCATION		Education of the respondent's spouse
NUMHH		Number of household members
INCOME		Annual income of the family
TOTINCOME		Total income
CHARITY		Charitable contributions
FACE		Amount that the company will pay in the event of the death of the named insured
FACECVLIFEPOLICIES		Face amount of life insurance policy with a cash value
CASHCVLIFEPOLICIES		Cash value of life insurance policy with a cash value
BORROWCVLIFEPOL		Amount borrowed on life insurance policy with a cash value
NETVALUE		Net amount at risk on life insurance policy with a cash value

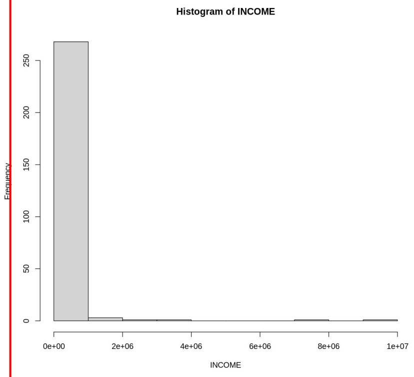
Source: Survey of Consumer Finances (SCF).

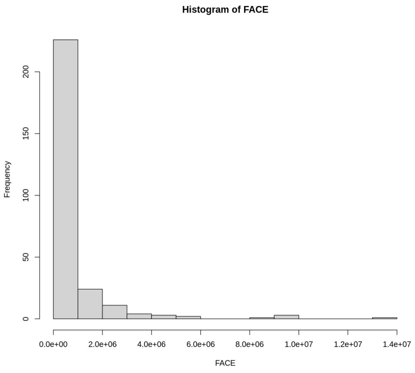
Data Overview





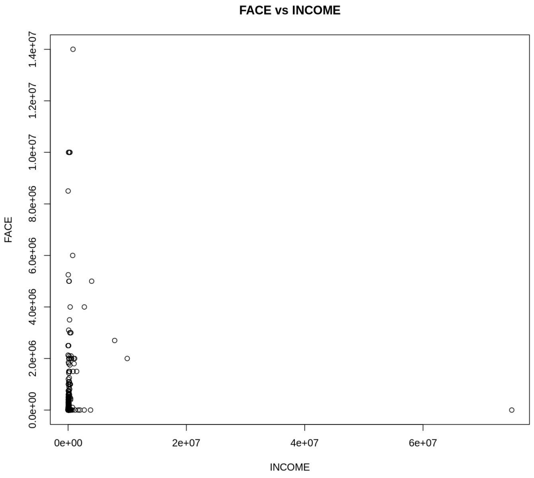
Highly skewed



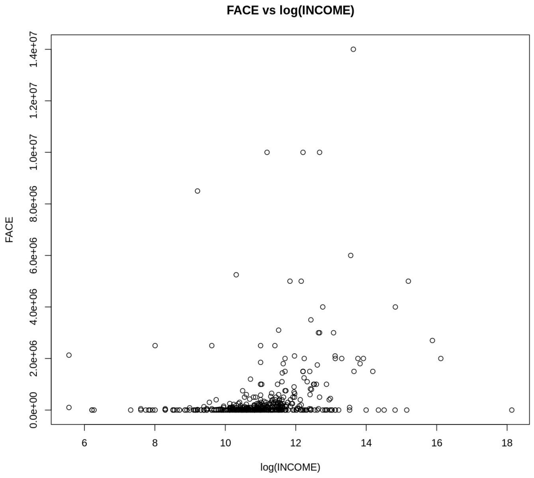



```
     GENDER               AGE               INCOME                FACE
 Min.   :0.000    Min.   :20.00    Min.   :      260    Min.   :
0
 1st Qu.:1.000    1st Qu.:37.00    1st Qu.:    28000    1st Qu.:
0
 Median :1.000    Median :47.00    Median :    54000    Median :
10000
 Mean   :0.826    Mean   :47.16    Mean   :   321022    Mean   :
411170
 3rd Qu.:1.000    3rd Qu.:58.00    3rd Qu.:   106000    3rd Qu.:
```

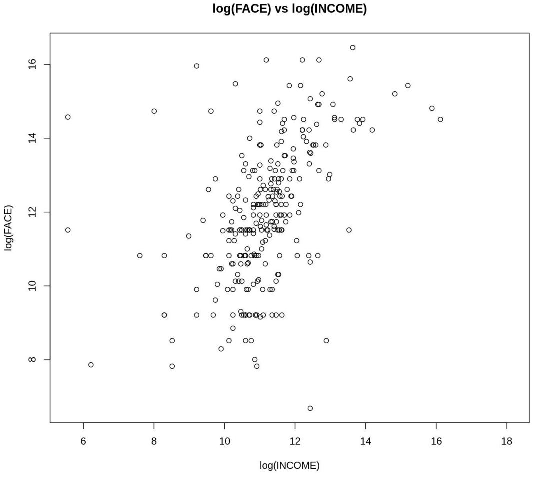
Dealing with Skewed Data



Face vs Income
Heavily clustered in bottom-left



Income Log Transformed
Still clustered around x-axis



Both Log Transformed
Much more useful for linear regression

Other Preparation

ZERO VALUES

- Numerous 0 values for FACE (45%)
- Cannot take log() of these
- Assumption: these value are where insurance was not purchased
- Therefore, we should drop these from our analysis

Our analysis will focus on the data we have on individuals that actually purchased insurance.

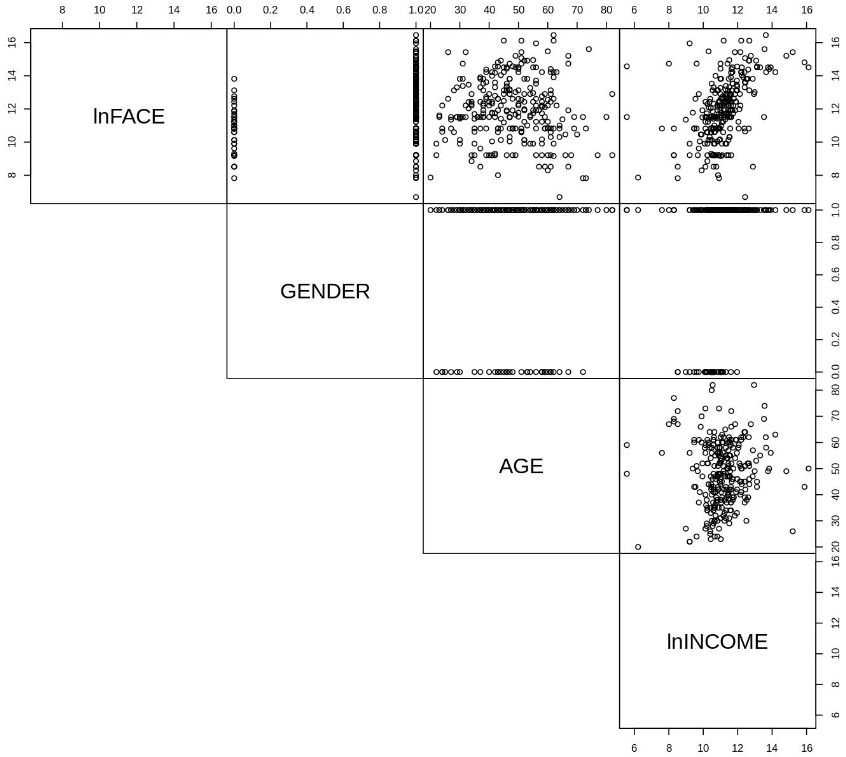
Exploring Linear Relationships

Pearson Correlation

GENDER	AGE	INCOME	lnINCOME
0.26	-0.05	0.25	0.48

Conclusions

- Age does **not** appear to be linearly correlated to lnFace (-0.05)
- Gender appears to have a **weak** linear relationship with lnFace (0.26)
- lnIncome appears to have a **strong** linear relationship with lnFace, relative to the others (0.48)



Data Modelling

Multiple Linear Regression

$$y = 4.50 + (0.876)x_1 + (-0.010)x_2 + (0.647)x_3$$

Interpretation

A value of 1 for **gender** leads to an increase in the predicted value for face by 140.0%.

For each unit increase in **age**, the model predicts a decrease in face by 1.01%

For a 1% increase in **income** the model predicts a 64.7% increase in face (the payout amount).

	Variable	Coefficient	Std. Err.
-	(Intercept)	4.501821	0.915055
X1	GENDER	0.875621	0.293576
X2	AGE	-0.010196	0.007952
X3	lnINCOME	0.647437	0.077576

Overall Performance

PERFORMANCE STATISTICS

Multiple R-squared: 0.261

Adjusted R-squared: 0.2528

Mean Squared Error (MSE): 2.577

Residual standard error (RSE): 1.617

F-statistic: 31.91 on 3 and 271 DF

p-value: < 2.2e-16

CONCLUSIONS

The model explains **25.3%** of the variance in lnFACE using the three explanatory variables.

Any predicted value using this model will have a 95% chance to be **+/- 3.23** of the line (*2 * RSE*).

Assessing Significance

Variable	T-Value	P-Value	Significant?
(Intercept)	4.920	1.51e-06	Yes
GENDER	2.983	0.00312	Yes
AGE	-1.282	0.20091	No
lnINCOME	8.346	3.65e-15	Yes

Assuming a confidence threshold of 95%, any p-value greater than 0.05 will be considered NOT statistically significant.

CONCLUSIONS

It was found that **gender** (p=0.003) and **income** (p=0.00) significantly predicted values for **face**.

Furthermore, **income** was found to be a stronger predictor than **gender**.

Lastly, **age** (p=0.201) was found to NOT be a significant predictor.

Summary of Results

Multiple linear regression was used to test if the age, gender and income of an individual that purchased life insurance would significantly predict the amount the company would pay in the event of death.

The fitted regression model was:

```
lnFACE = 4.50 + (0.876)GENDER + (-0.010)AGE + (0.647)lnINCOME
```

The overall regression was statistically significant ($R2 = 0.261$, $F(3, 271) = 31.91$, $p < 2.2e\text{-}16$).

Age was found to <u>not</u> be a good predictor of face ($p=0.201$).

Gender was found to significantly predict face ($p=0.003$). A value of 1 for gender, if all other variables remain fixed, leads to an increase in the predicted value for face by 140.0%. (It is unknown if this value for gender implies male or female for this data set.)

Income was found to significantly predict face ($p=0.000$). For a 1% increase in income, if all other variables remain fixed, the model predicts a 64.7% increase in face (the payout amount).

Thank You!

Colin Bowers

colin.bowers@gmail.com

https://www.linkedin.com/in/colinbowers/

https://github.com/straylight77