Linear Regression Examples

Areej Alasiry

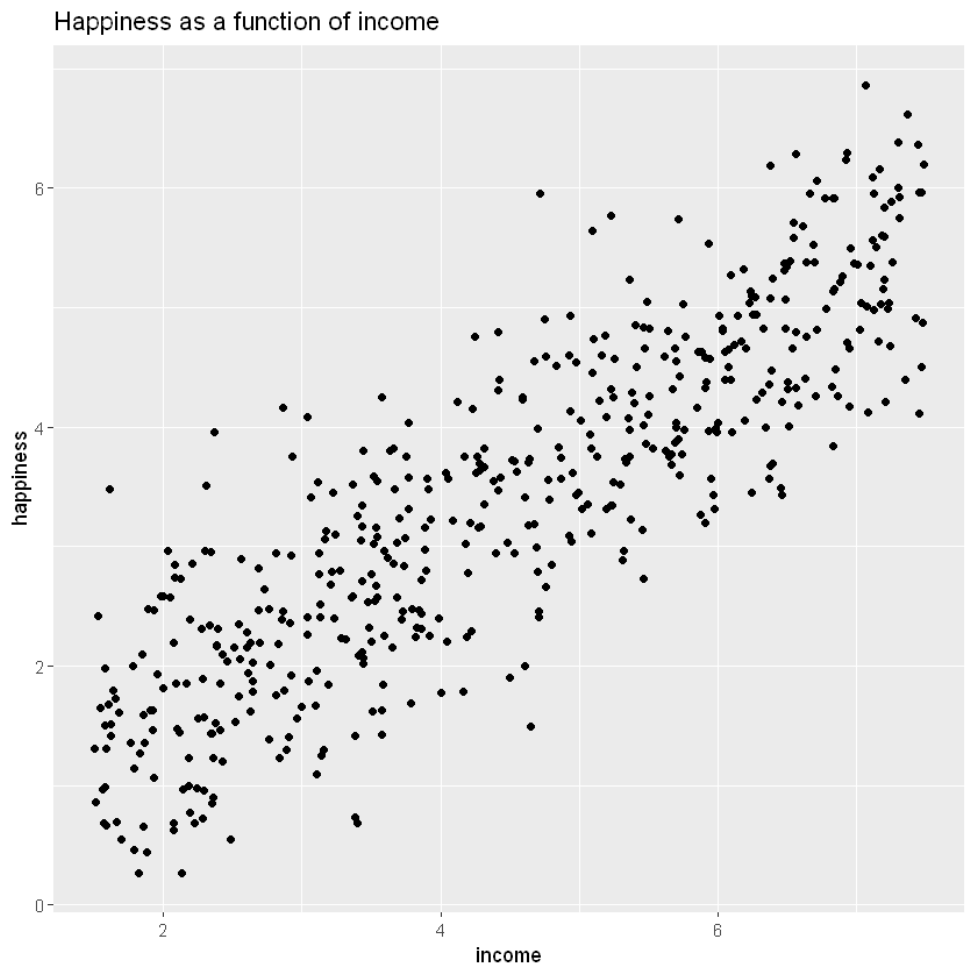
Example 1
Income and Happiness

Road Map

- Explore the dataset
- Split the dataset into training and test, randomly (50-50 split)
- Use the training dataset to fit the model
- Use the test dataset to evaluate the model
- Interpret the results and extract knowledge/advice.

Income Happiness Dataset

income	happiness
3.862647	2.314489
4.979381	3.433490
4.923957	4.599373
3.214372	2.791114
7.196409	5.596398
3.729643	2.458556

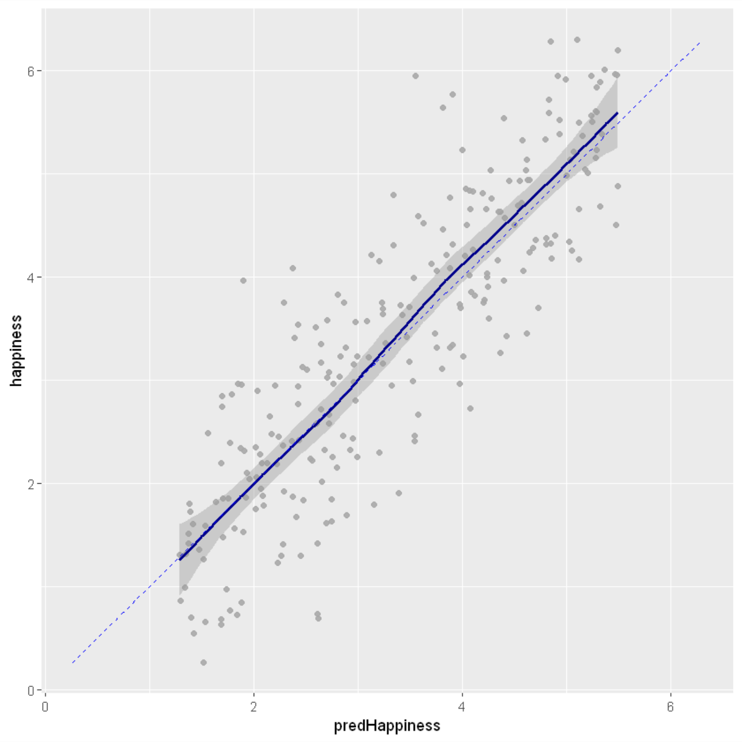


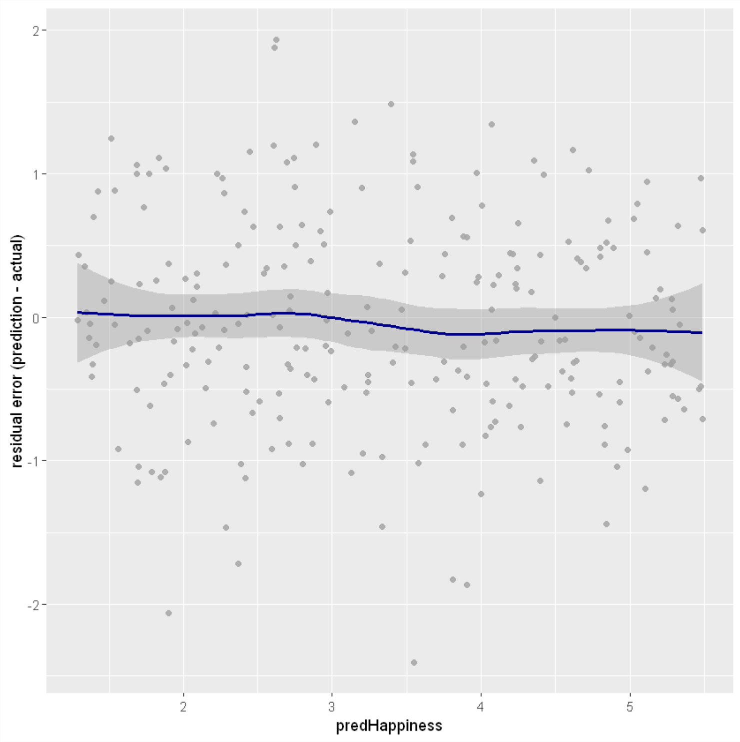
Residuals

- Prediction Quality
  - Systematic errors?

Analyse the Residuals

```
Residuals:
Min         1Q          Median      3Q          Max
-1.99990    -0.47966    -0.01526    0.48223     2.11681
```





R-Squared

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(p - o)^2}{\sum_{i=1}^{n}(\mu - o)^2}$$

- Over the training dataset: 0.7445
- Over the test dataset: 0.7531

Interpreting Coefficients

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.23226    0.12615   1.841   0.0668 .
income       0.70245    0.02613  26.881   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The interpretation of the slope (value = 0.70245) is that happiness rate increases 0.70 units, on average, for each one unit (one percent) increase in the income.
- The interpretation of the intercept (value=0.23226) is that if income = 0, the predicted average happiness rate would be 0.23

Example 2
Predict Income

Linear Regression

Suppose you want to predict personal income of any individual in the general public, within some relative percent, given their age, education, and other demographic variables. In addition to predicting income, you also have a secondary goal: to determine the effect of a bachelor's degree on income, relative to having no degree at all.

- *From: Practical Data Science with R, 2nd Edition. Nina Zumel and John Mount. [Chapter 7]*

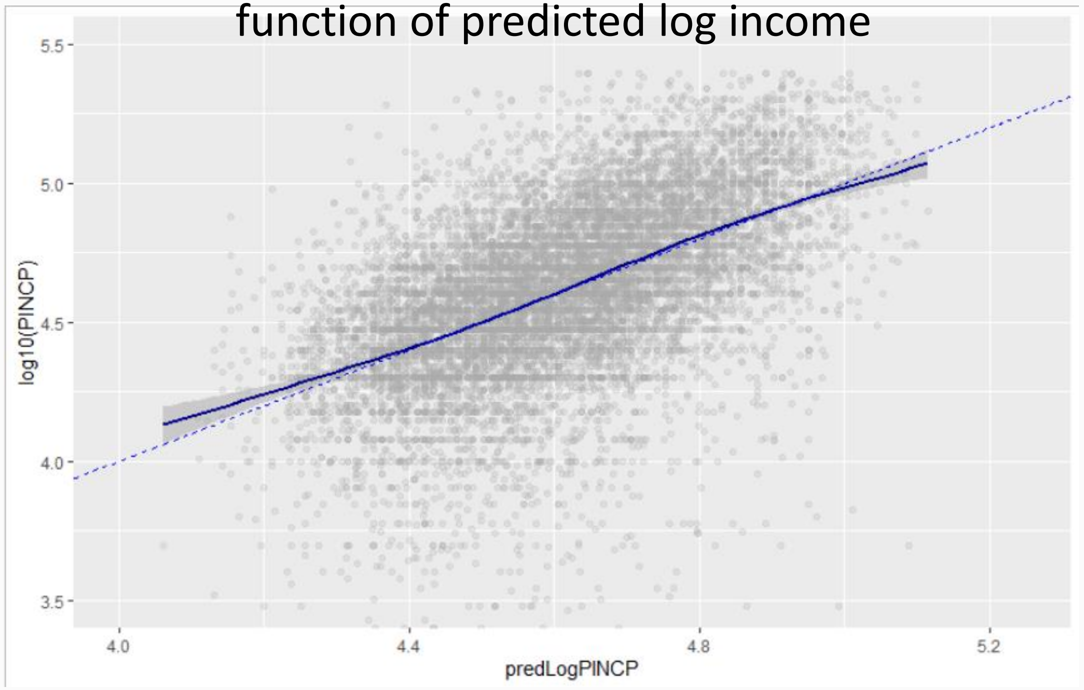
The dataset

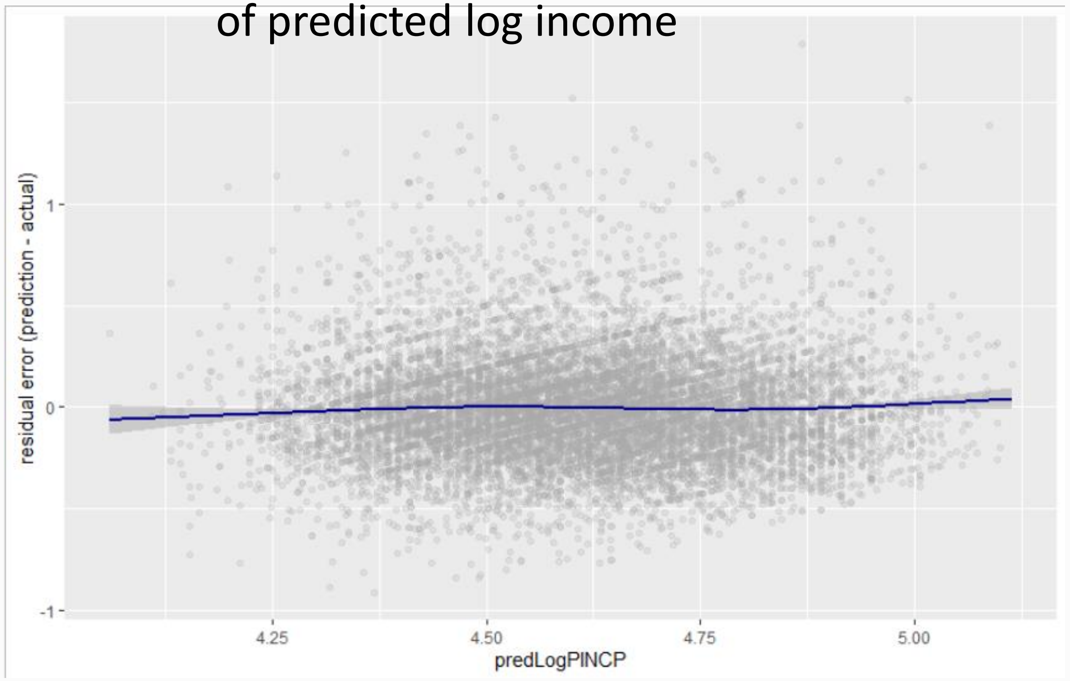
- Records = 22241
- Features = 204

Formula: log10(PINCP) ~ AGEP + SEX + COW + SCHL

RT	SERIALNO	SPORDER	PUMA	ST	ADJINC	AGEP	CIT	CITWP	COW	...	FSEXP	FSSIP	FSSP	FWAGP	FWKHP	FWKLP
P	000006646	03	02400	Alabama/AL	1007588	24	Born in the U.S.	NA	Employee of a private for profit	...	No	Yes	Yes	Yes	Yes	Yes
P	000008359	04	02702	Alabama/AL	1007588	31	Born in the U.S.	NA	Private not-for-profit employee	...	No	No	No	Yes	No	No
P	000015018	01	00400	Alabama/AL	1007588	26	Born in the U.S.	NA	Employee of a private for profit	...	No	No	No	No	No	No
P	000017383	04	02400	Alabama/AL	1007588	27	Born in the U.S.	NA	Employee of a private for profit	...	No	No	No	No	No	No
P	000030038	02	02100	Alabama/AL	1007588	27	Born in the U.S.	NA	Private not-for-profit employee	...	No	No	No	No	No	No
P	000033559	02	02500	Alabama/AL	1007588	47	Born in the U.S.	NA	Employee of a private for profit	...	No	No	No	No	No	No

Analysing the Residuals





Extract Relations and Knowledge

- Examples:

```
Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         4.0058856  0.0144265 277.676  < 2e-16 ***
AGEP                                0.0115985  0.0003032  38.259  < 2e-16 ***
SEXFemale                          -0.1076883  0.0052567 -20.486  < 2e-16 ***
COWFederal government employee      0.0638672  0.0157521   4.055 5.06e-05 ***
COWLocal government employee       -0.0297093  0.0107370  -2.767 0.005667 ** 
COWPrivate not-for-profit employee -0.0330196  0.0102449  -3.223 0.001272 ** 
COWSelf employed incorporated       0.0145475  0.0164742   0.883 0.377232    
COWSelf employed not incorporated  -0.1282285  0.0134708  -9.519  < 2e-16 ***
COWState government employee       -0.0479571  0.0123275  -3.890 0.000101 ***
SCHLRegular high school diploma     0.1135386  0.0107236  10.588  < 2e-16 ***
SCHLGED or alternative credential   0.1216670  0.0173038   7.031 2.17e-12 ***
SCHLsome college credit, no degree  0.1838278  0.0106461  17.267  < 2e-16 ***
SCHLAssociate's degree              0.2387045  0.0123568  19.318  < 2e-16 ***
SCHLBachelor's degree               0.3637114  0.0105810  34.374  < 2e-16 ***
SCHLMaster's degree                 0.4445777  0.0127100  34.978  < 2e-16 ***
SCHLProfessional degree             0.5111167  0.0201800  25.328  < 2e-16 ***
SCHLDoctorate degree                0.4818700  0.0245162  19.655  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residuals Summary

- Over the training dataset

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.5038	-0.1354	0.0187	0.0000	0.1710	0.9741

- Over the test dataset

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.789150	-0.130733	0.027413	0.006359	0.175847	0.912646

Linear Regression

- Linear regression assumes that the outcome is a linear combination of the input variables.
- If you want to use the coefficients of your model for advice, you should only trust the coefficients that appear statistically significant
- Overly large coefficient magnitudes, overly large standard errors on the coefficient estimates, and the wrong sign on a coefficient could be indications of correlated inputs.
- Linear regression can predict well even in the presence of correlated variables, but correlated variables lower the quality of the advice.
- Linear regression will have trouble with problems that have a very large number of variables, or categorical variables with a very large number of levels.

Try the following

- Measure R-square for the following:
  - O = (1,2,3,4,5,9,10)
  - P = (0.5,0.5,0.5,0.5,0.5,9,10)

Reference

- *Practical Data Science with R, 2nd Edition. Nina Zumel and John Mount. [Chapter 7]*