Simple Linear Regression, Inference

CSU Chico, Math 314

2018-11-28

outline

outline

Recap, linear regression

Linear regression is a method to fit a line through a scatter plot of data in a "best" sense. Often, interest lies in the relationship between the explanatory and the response variable.

outline

Inference, β_0, β_1

Linear regression estimates the population parameters β_0 (intercept) and β_1 (slope), just like every other parameter we have estimated. As such, the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ of these parameters have their own sampling distributions.

Inference, β_0, β_1

It turns out that the estimators $\hat{\beta}$ are also approximately normally distributed when the sample size is sufficiently large,

$$\frac{\hat{\beta} - \beta}{\sigma_{\hat{\beta}}} \sim N(0, 1).$$

Inference, hypothesis tests

Hypothesis testing naturally follows. The most common hypothesis test for linear regression parameters is

$$\begin{aligned} H_0: &\quad \beta = 0 \\ H_A: &\quad \beta \neq 0. \end{aligned}$$

with $\alpha = 0.05$.

Inference, a note

A small but often missed point. Hypothesis testing is meaningless if the model is wrong.

George E. P. Box

Essentially, all models are wrong, but some are useful.

Standard Output

The hypothesis test above has a natural and informative interpretation in most contexts.

```
url <- "https://roualdes.us/data/elmhurst.csv"
elmhurst <- read.csv(url)
elmReg <- lm(gift_aid ~ family_income, data=elmhurst)
# summary(elmReg) # RStudio
```

Standard Output

Be sure to understand and be able to find

- standard errors
- t values
- p-values
- adjusted R^2

Inference, confidence intervals

If we can do hypothesis testing, we can do confidence intervals. The function `confint` in `R` is extremely helpful.

```
# use lm fitted model, elmReg from above
confint(elmReg) # default is 95%

##                      2.5 %      97.5 %
## (Intercept)   21.72269421 26.91596380
## family_income -0.06480555 -0.02133775

confint(elmReg, level=0.99) # can specify confidence

##                      0.5 %      99.5 %
## (Intercept)   20.85539590 27.78326212
## family_income -0.07206486 -0.01407844
```

outline

R^2

It is common to use the square of the (Pearson) correlation to explain the strength of a linear fit.

R^2

The R^2 of a linear model describes the amount of variation in the response variable y that is explained by the least squares line on the explanatory variable x.

R^2, example

Using the data frame `elmhurst`, the correlation between gift aid and family income is $R = -0.4986$. Thus, $R^2 = 0.2486$. We say 24.86% of the variation in `gift_aid` is explained by the least squares line on `family_income`.

Adjusted R^2

It is almost[1] always better to use the "Adjusted R-squared" value that R gives you in all linear regression output. The problem with the R^2 above is that the more explanatory variables you use the higher this value is – this doesn't always conform to reality.

[1] Except when the adjusted R^2 is negative.

Adjusted R^2

adjusted R^2

The adjusted R^2 of a linear model describes the amount of variation in the response variable y that is explained by the least squares line on the explanatory variable x, after taking into account the number of explanatory variables.

outline

Extrapolation, example

At age 8, Shaquille O'Neal was 4'8". At age 16, he was 6'8". Can we use these data to predict how tall Shaq is now that he is 43? In eight years, Shaq grew $2'$. 27 years later, Shaq should be $6'9"$ taller than he was at 16, thus $13'7"$. Sound reasonable?

Extrapolation

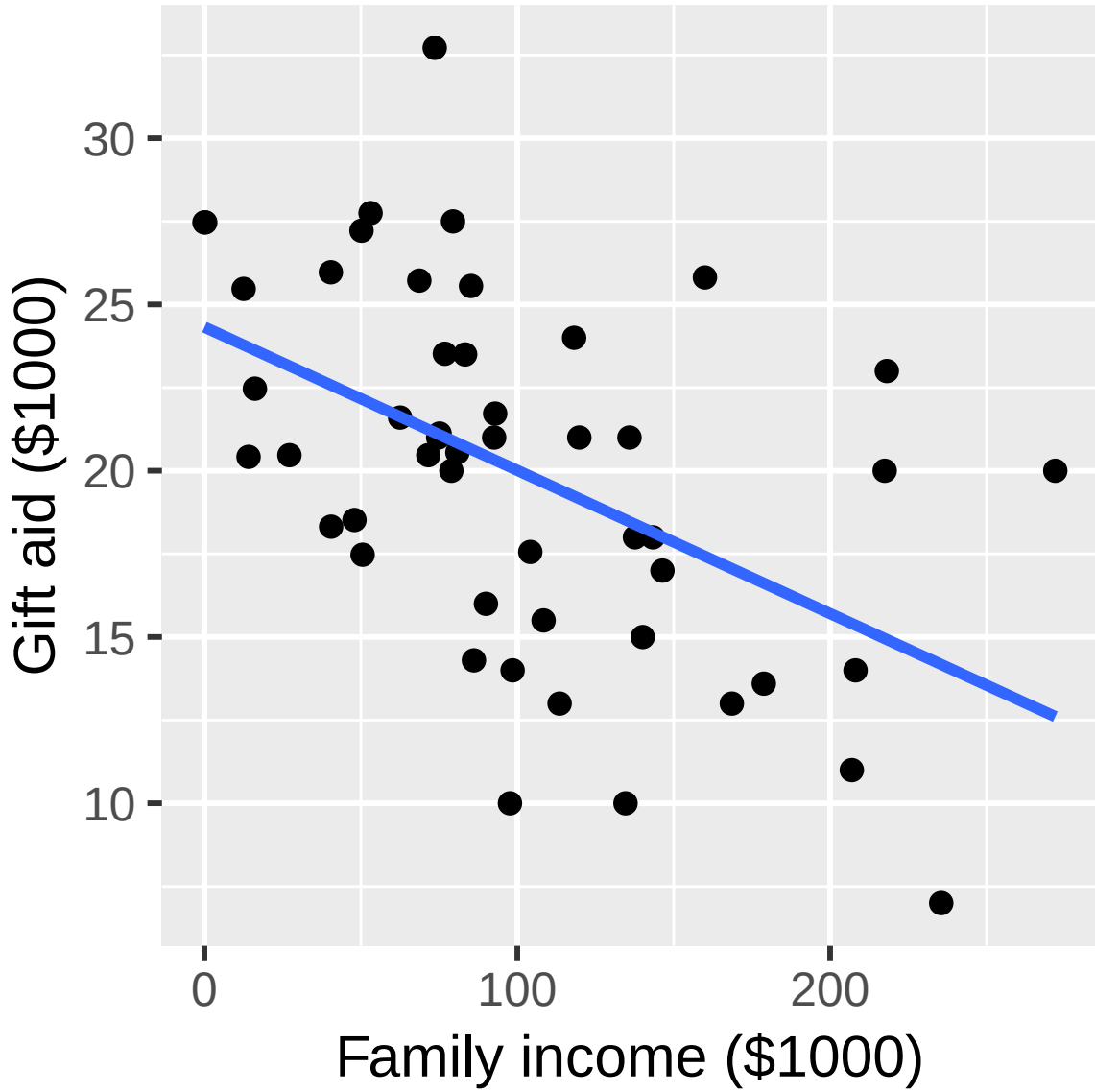
Extrapolation is in general dangerous. Sometimes it works, but not often so watch out.

extrapolation

Applying a model to values outside of the range of the original data is called extrapolation.

Extrapolation, example

How much gift aid would a student expect to receive if their family income was $1 million?



Extrapolation, example

How much would gift aid would a student expect to receive if their family income[2] was \$1 million? Using our least squares line,

$$\widehat{aid} = 24.32 + -0.04 \times family\_income$$

we'd estimate -18.75 thousand dollars.

```
?predict.lm
```

[2] Don't forget family income is in units of \$1000.

outline

Take away

- Hypothesis testing and confidence intervals live on
- most linear regression software output defaults to two-sided alternatives
- Too many people rely on R^2, use adjusted R^2 instead.
- Extrapolation is dangerous – be careful.

outline

references I

David M Diez, Christopher D Barr, and Mine Cetinkaya-Rundel. *OpenIntro Statistics*. CreateSpace independent publishing platform, third edition, 2015.

Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer Science and Business Media, 2009.