Linear Regression: Intercept/Slope Transformations & Regression Diagnostics

Andrew Stokes

April 18, 2017

Review of epidemiological concepts

Recall the two types of error in epidemiological analyses:

- Random error
- Systematic error

The first of these occurs due to chance, often as a result of small sample sizes. The second is a result of biases. While the first can often be addressed by collecting more data, the second type is more difficult to address and often requires statistical adjustment.

Biases in epidemiological studies fall into three general categories:

- selection bias
- information bias
- confounding

Selection Bias

Let me give an example of selection bias from my own research. Suppose we ask how does a history of obesity affect your wellbeing late in life?

To answer this question, we use cross-sectional data drawn from the National Health and Nutrition Examination Survey (NHANES) on non-institutionalized adults ages 75 and older in the United States.

We decide to assess wellbeing using self-reported health, assessed on a Likert-scale. Unfortunately, the individuals' obesity histories were not directly observed, but we know their weight at age 25 based on a question in the NHANES that asked them to recall this information.

We decide that we'll analyze the data using a logistic regression model. The dependent variable is a dummy of poor self-rated health and the key independent variable is a continuous measure of body mass index at age 25. This variable is calculated by combining recalled weight at age 25 with height reported at the time of survey.

What sources of *selection bias* should we be aware of when investigating this association? Let's focus for now on this one type of bias and not the other types, which may also be present.

Another type of selection bias is *non-response* bias. Household surveys often ask about smoking. Heavy smokers, the target of the study are less likely to respond.

Examples from your projects?

Information Bias

To illustrate information bias, let's return to the example above on weight history and self-reported health among older adults. This study design is subject to at least two different sources of information bias! Can you identify them?

Confounding

Confounding is when the association between exposure and outcome is affected by some extraneous factor. To be a confounder, a variable must be:

- associated with the exposure
- associated with the disease
- not in the causal pathway between exposure and disease

Let's take the example of coffee drinking (exposure) and heart disease (outcome) and consider confounding by smoking.

- smoking is related to coffee drinking (positively)
- smoking is also related to heart disease (positively)
- smoking is not on the causal pathway between coffee drinking and heart disease

What's the direction of the bias (positive or negative)? In this case, confounding by smoking would lead to an overestimation of the true effect size. Since smoking is both positively related to coffee drinking and heart disease, it will create the appearance of a stonger relationship between these variables relative to the true association.

Confounding can work in the other direction too. Take the following example, where obesity is the independent variable and mortality is the dependent variable. Let's again consider confounding by smoking.

Smoking is negatively related to obesity. That is, smokers tend to be leaner than non-smokers. This is in part because its an appetite suppresent. It also may be a matter of behavioral substitution.

Smoking is also strongly positively related to mortality. That is the risks of death among smokers is much greater than for non-smokers.

As a result of the negative association between smoking and obesity and the positive association between smoking and mortality, if smoking is ignored, it will appear that obesity lowers risks of dying. That obesity is good for you.

Sadly, this is not just a hypothetical example. Many research studies have come to this conclusion due to a failure to adjust for confounding by smoking as well as other biases.

Conceptualizing associations between variables

Let's recall the different ways in which variable can be related to each other. It will be useful to use visual tools such as directed acyclic graphs (DAGS) to conceptualize relationships between variables in your study. Doing so will help you make sure you are pursuing an appropriate analytic strategy.

- direct
- indirect
- effect modification
- confounding

Let's examine what each of these look like visually and consider some examples.

Hill's criteria for causation

Is the assocation in your study causal? Let's consider Hill's classic criteria:

- temporal relationship
- strength of the association
- dose-response relationship
- consistency of the association
- specificity of the association
- biological plausibility
- coherence with existing knowledge
- experimental evidence
- analogy

In practice

In practice, as an epidemiologic researcher:

- Careful review and judgment of all relevant information available
- Look for possibility of random and systematic errors
- Judge whether an observed association is spurious, indirect or real.

Linear regression (Board Work)

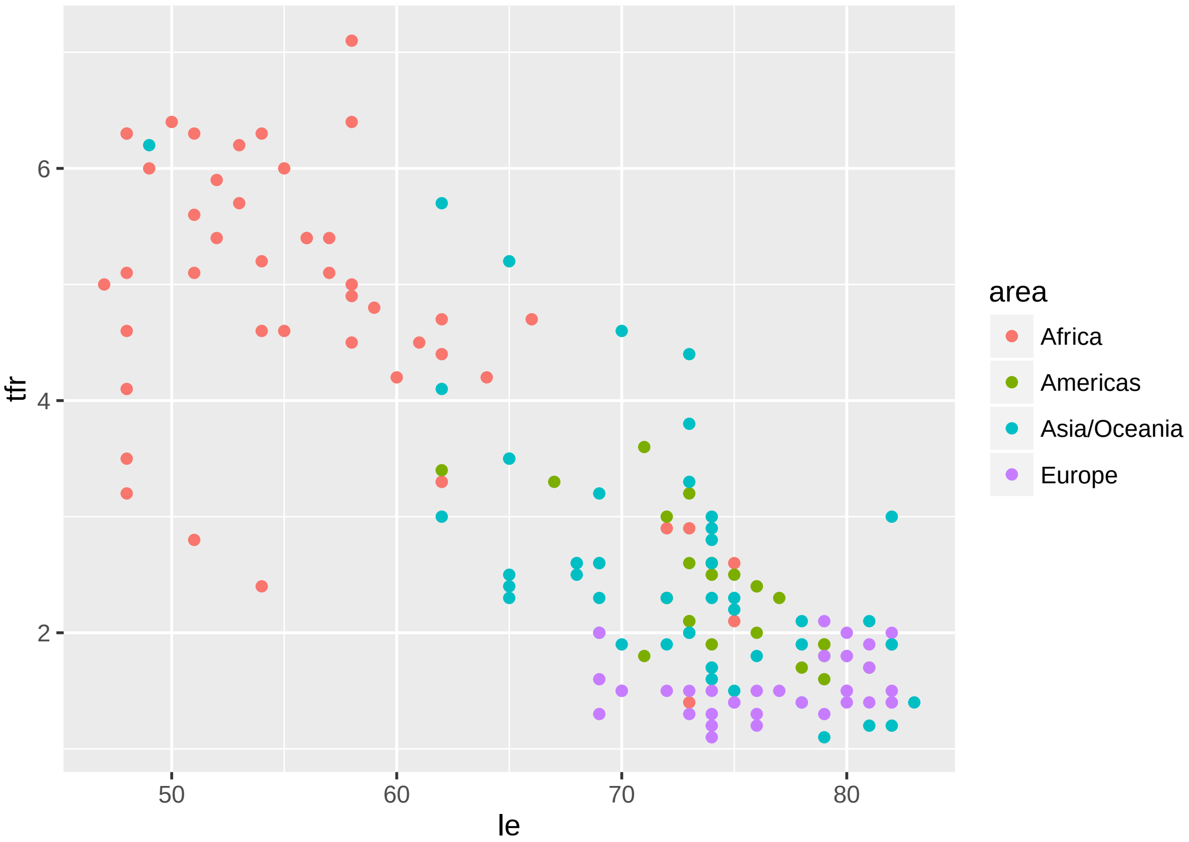
Demonstration

Get the ggplot2 library and read in the dataset

```
library("ggplot2")
setwd("C:/Users/acstokes/Desktop/GH 811/Fall 2016/Week 11/ggplot2workshopFiles")
w <- read.csv(file="WDS2012.csv", head=TRUE, sep=",")
```

Visualize the relationship between life expectancy and the total fertility rate

```
p <- ggplot(data=w, aes(x=le, y=tfr, color=area))
p + geom_point()
```



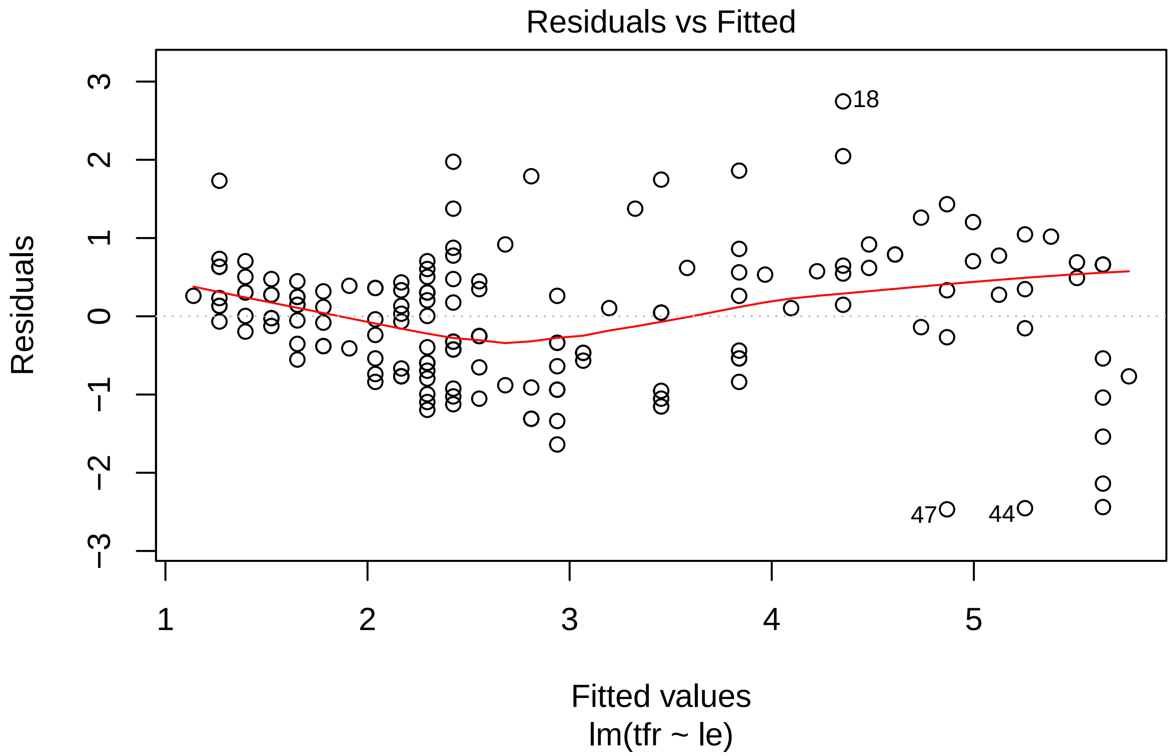
Basic model

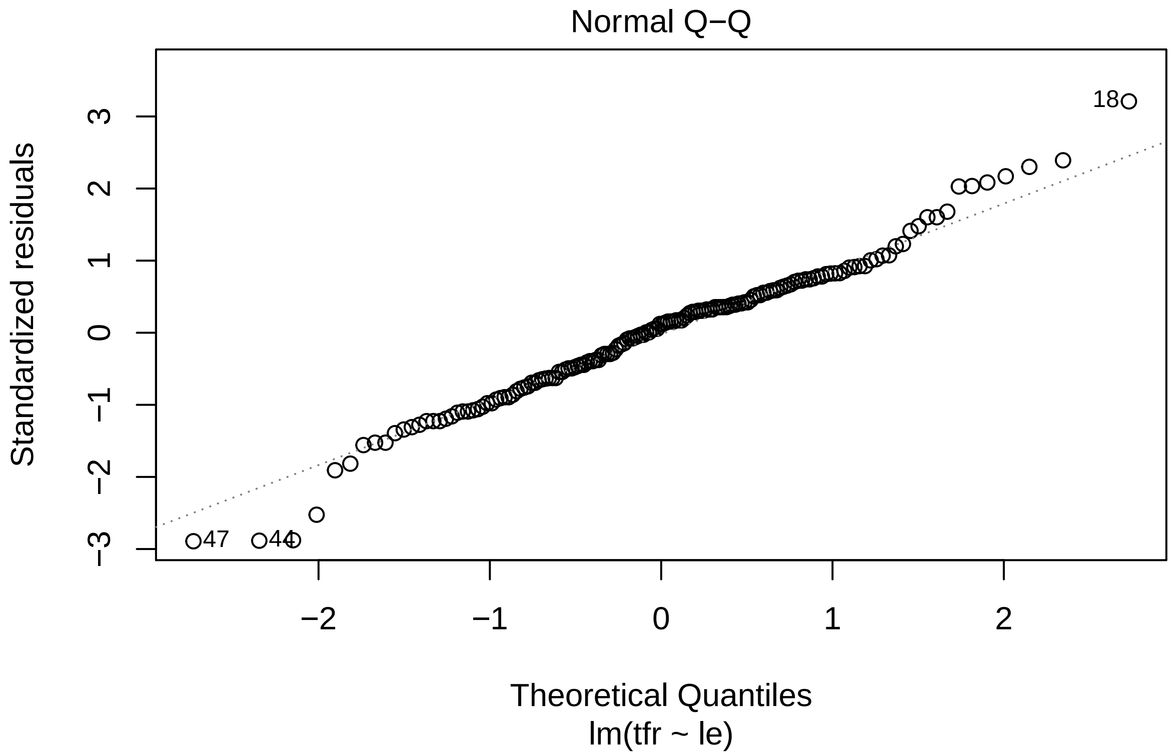
```
fit <- lm(tfr ~ le, data = w)
fit
```

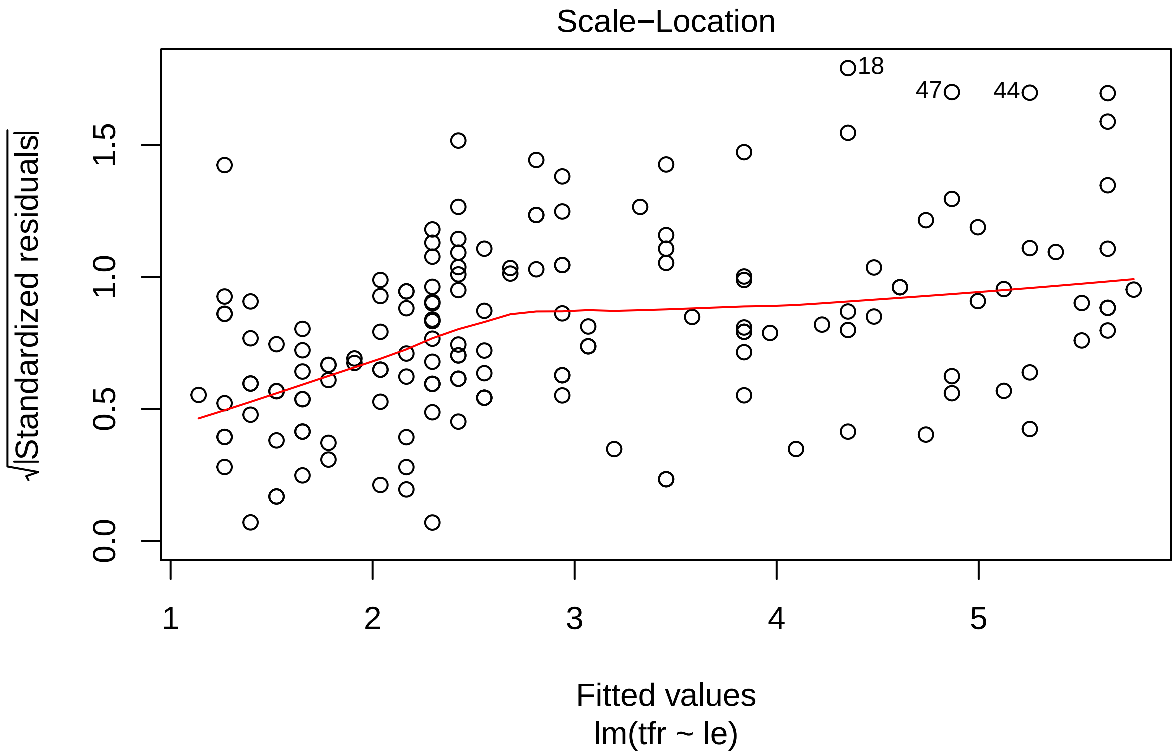
```
##
## Call:
## lm(formula = tfr ~ le, data = w)
##
## Coefficients:
## (Intercept)           le
##     11.8106      -0.1286
```

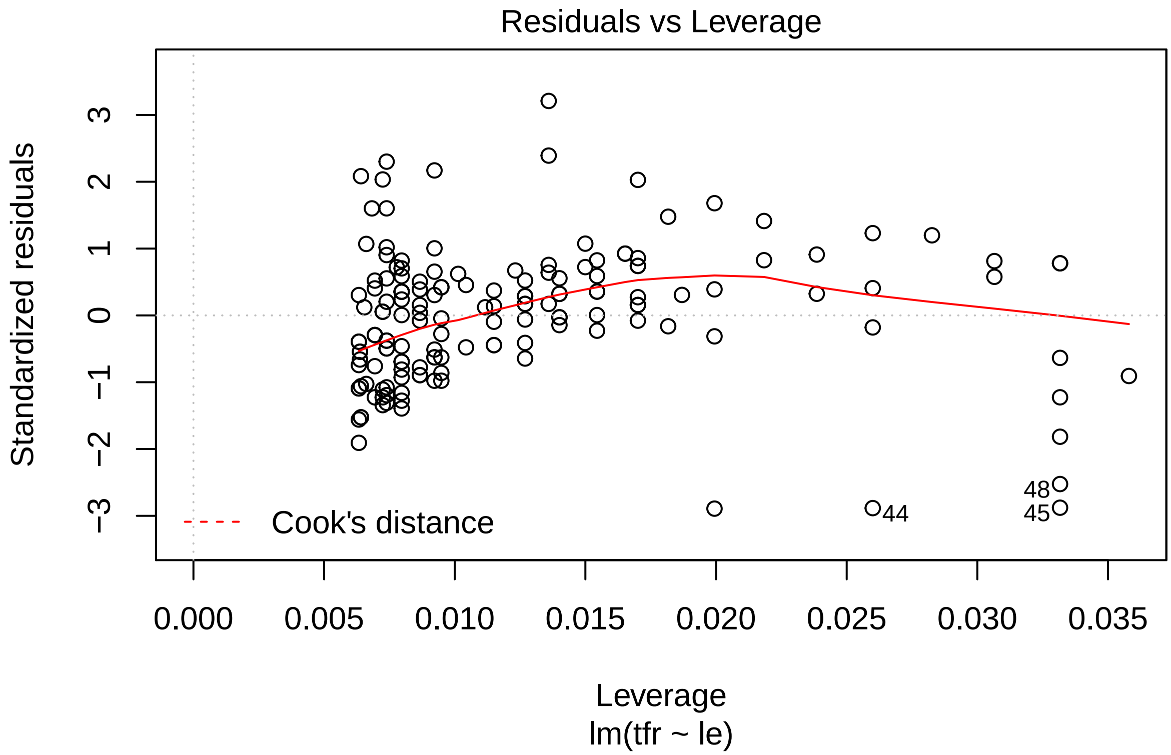
Model diagnostics

```
plot(fit)
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
18
47
44
−2
−1
0
1
2
Theoretical Quantiles
lm(tfr ~ le)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1
2
3
4
5
18
47
44
Fitted values
lm(tfr ~ le)



Model with more meaningful intercept

```
fit2 <- lm(tfr ~ I(le - mean(le)), data = w)
fit2
```

```
##
## Call:
## lm(formula = tfr ~ I(le - mean(le)), data = w)
##
## Coefficients:
##      (Intercept)  I(le - mean(le))
##           2.9582           -0.1286
```

Effect per 10 unit increase in life expectancy

```
fit3 <- lm(tfr ~ I(le * (1/10)), data = w)
```

Let's make some predictions bsaed on the model. What is the expected value of the TFR at life expectancies of 50, 60, 70, 80 and 83?

```
newx <- c(50, 60, 70, 80, 83)
coef(fit)[1] + coef(fit)[2] * newx
```

```
## [1] 5.381688 4.095903 2.810118 1.524334 1.138598
```

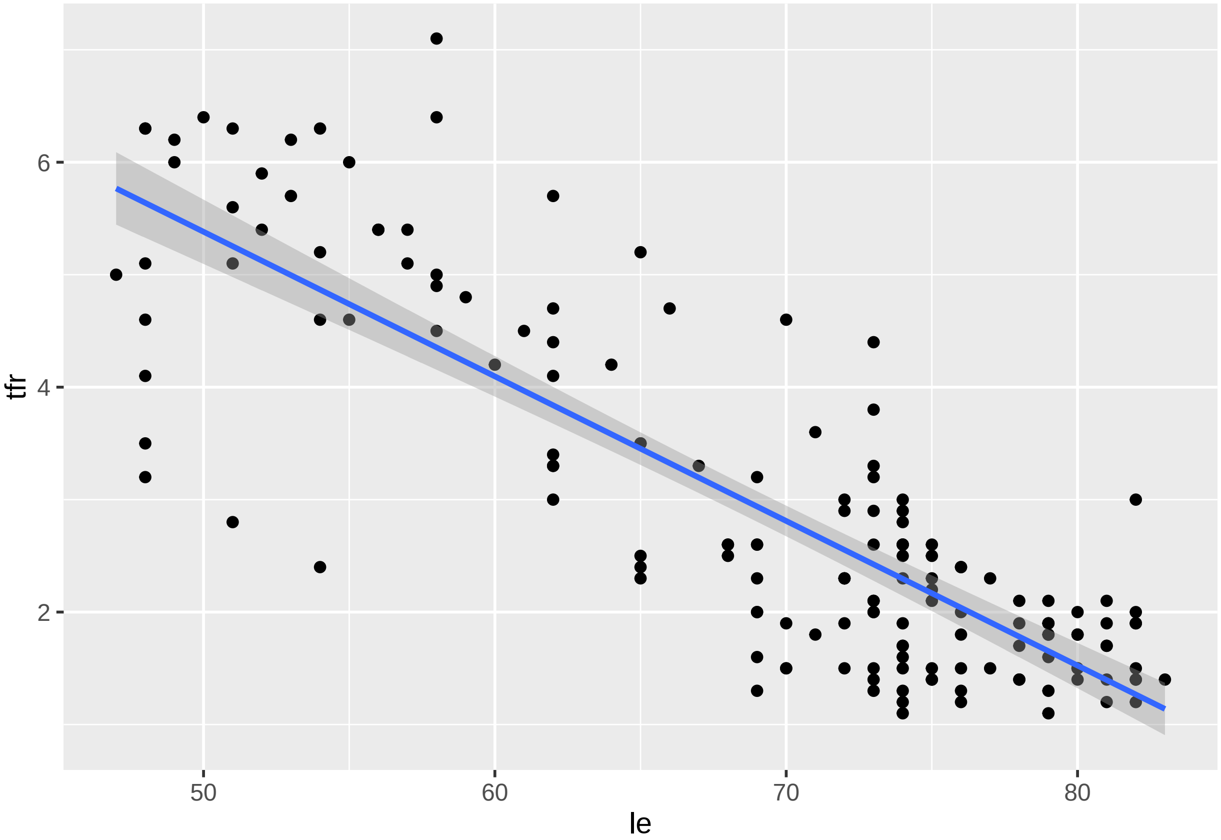
There is an easier way to implement the prediction.

```
predict(fit, newdata = data.frame(le = newx))
```

```
##        1         2         3         4         5
## 5.381688 4.095903 2.810118 1.524334 1.138598
```

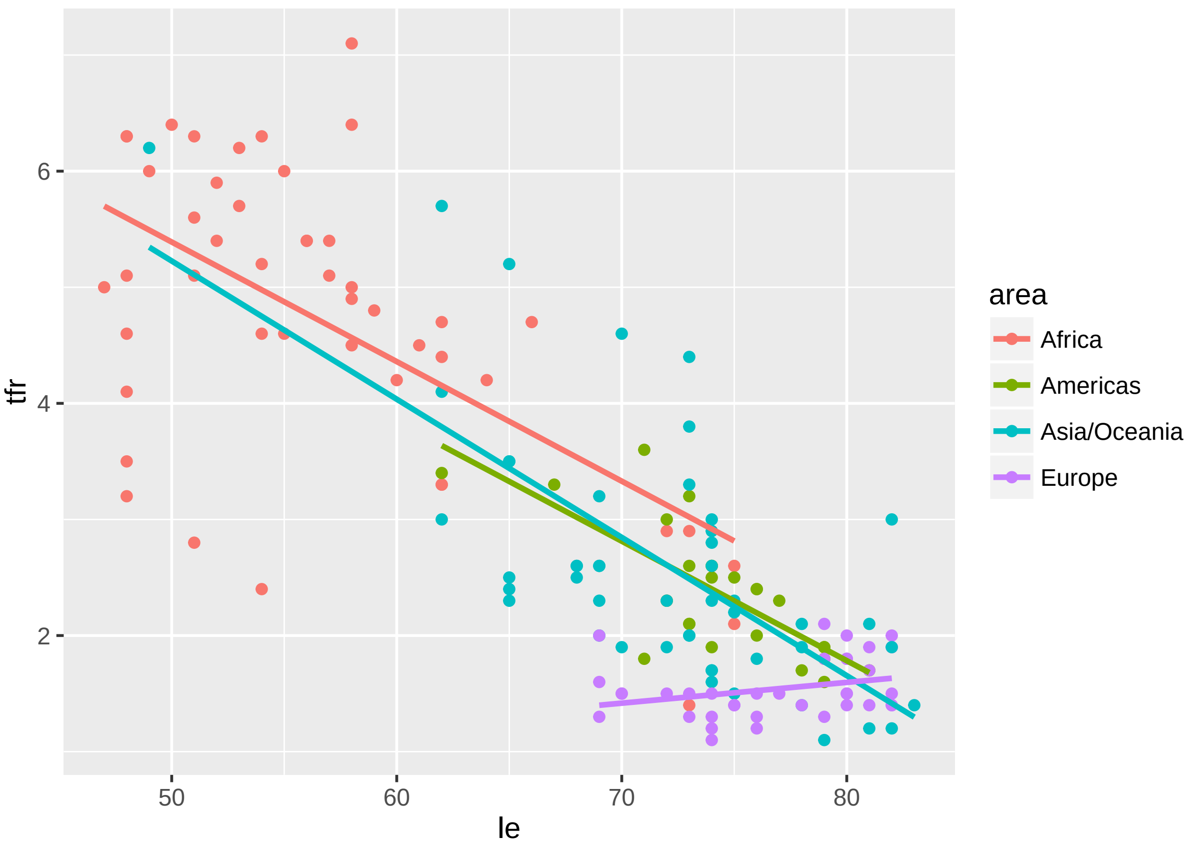
How does the slope look graphically?

```
p <- ggplot(data=w, aes(x=le, y=tfr))
p + geom_point() + geom_smooth(method="lm")
```



How about if we wanted to see how the slope differs by area of world? This is a visual way to explore interaction effects.

```
p <- ggplot(data=w, aes(x=le, y=tfr, color=area))
p + geom_point() + geom_smooth(method="lm", se=FALSE)
```



Sources:

- Caffo Linear regression text
- Princeton ggplot2 seminar
- IHME