Stat 411/511

Simple Linear Regression

Nov 18 2015

Charlotte Wickham

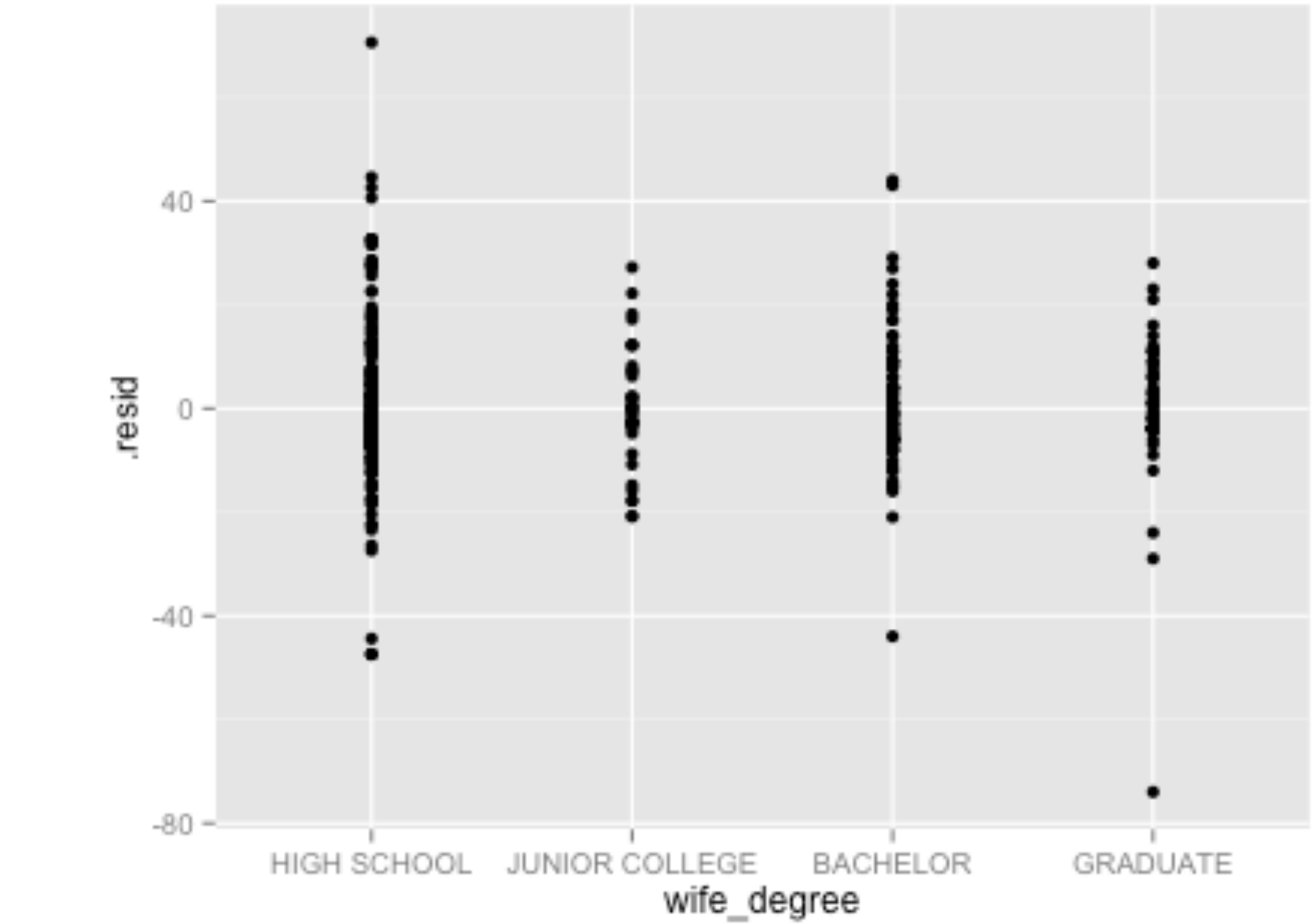
stat511.cwick.co.nz

DA#2

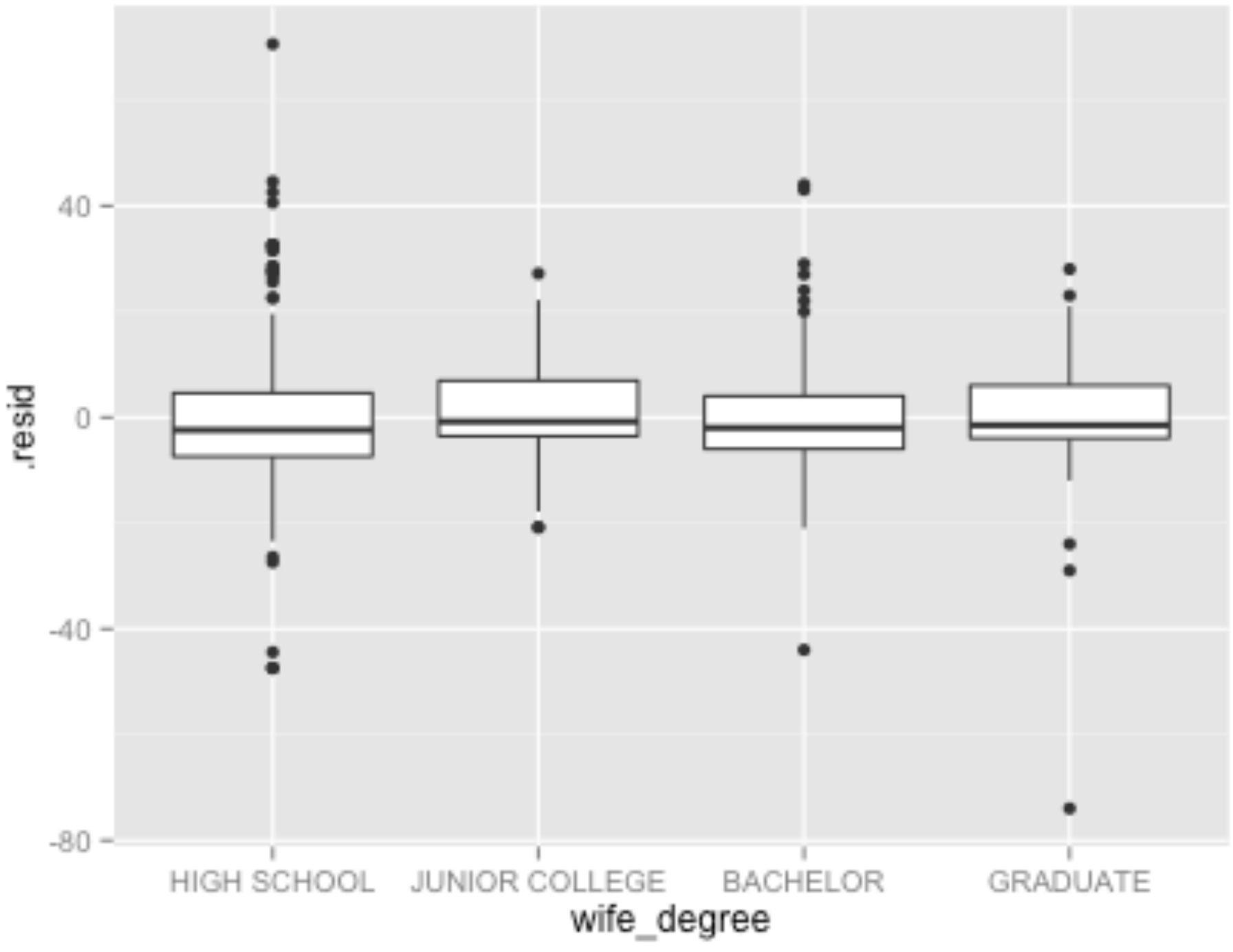
Q2 You should be reporting test results and confidence intervals. (No adjustments required)

Assumptions: don’t be fooled by different sample sizes. (add pic)

```
qplot(wife_degree, .resid,  data = fit)
```

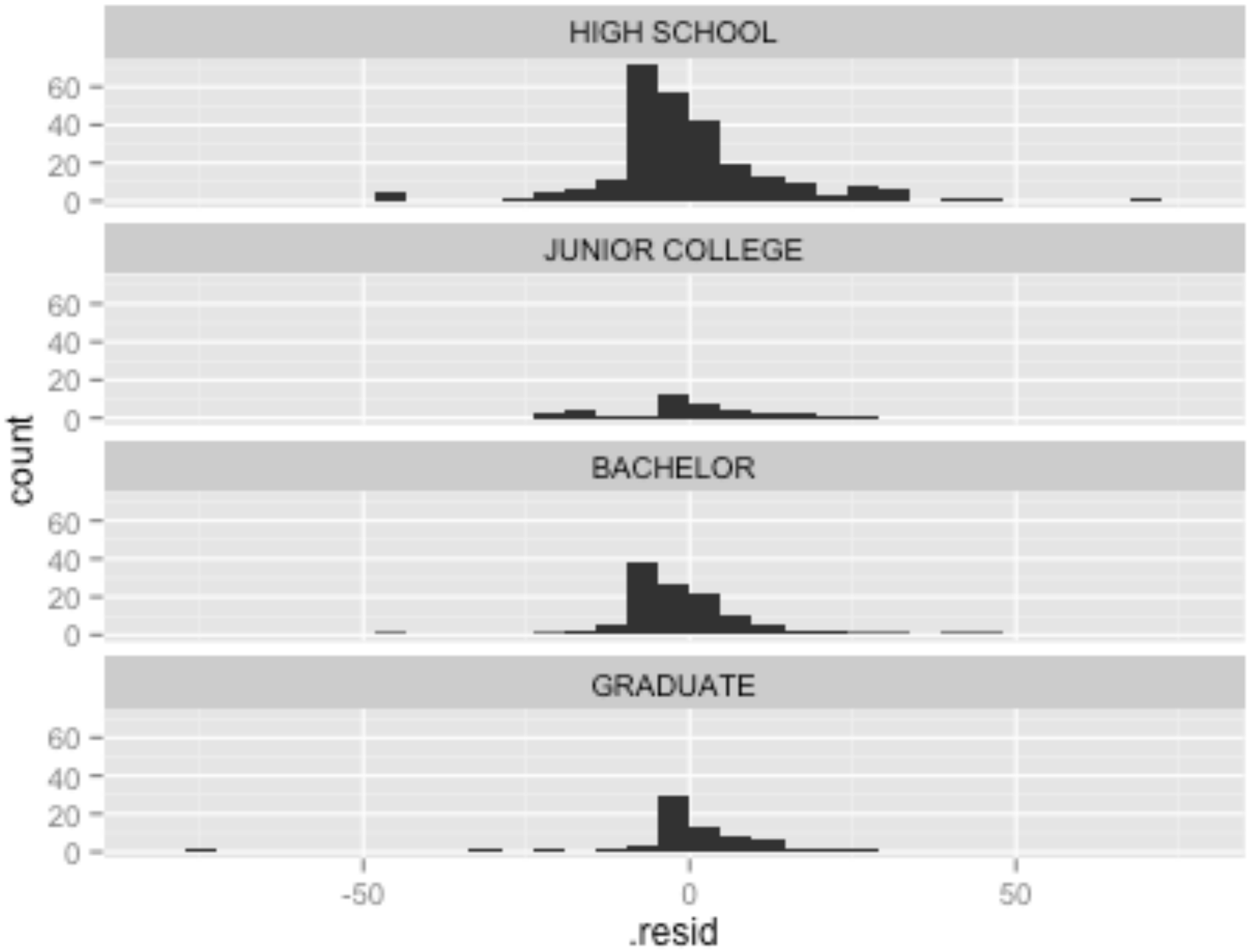


```
qplot(wife_degree, .resid,  data = fit,
  geom = "boxplot")
```

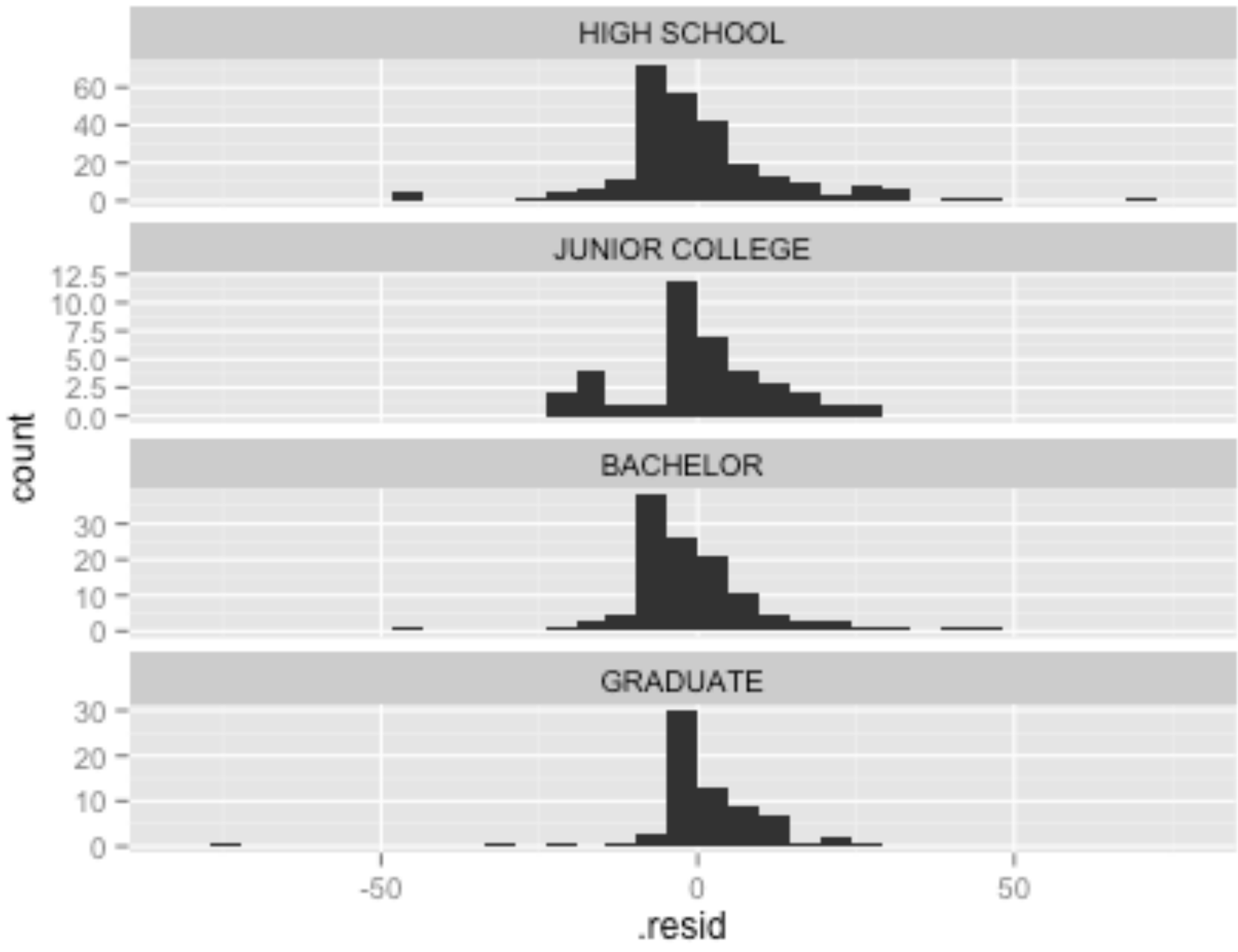



```
qplot(.fitted, .resid,  data = fit,
  geom = “boxplot”, group = wife_degree)
```

```
qplot(.resid,  data = fit, )  +
  facet_wrap(~ wife_degree, ncol = 1)
```



```
qplot(.resid,  data = fit, )  +
  facet_wrap(~ wife_degree, ncol = 1, scale = "free_y")
```



The **response** variable is the measurement we are interested in explaining or predicting.

The **explanatory** variable is the measurement we want to use to explain or predict the response.

So far...

We have had:

a single **continuous** response

quantitative

a single **discrete** explanatory variable

(the grouping variable)

qualitative
categorical

We have been interested in predicting the **mean response** in **each group.**

Simple Linear Regression

We have had:

a single **continuous** response

continuous

a single ~~**discrete**~~ explanatory variable

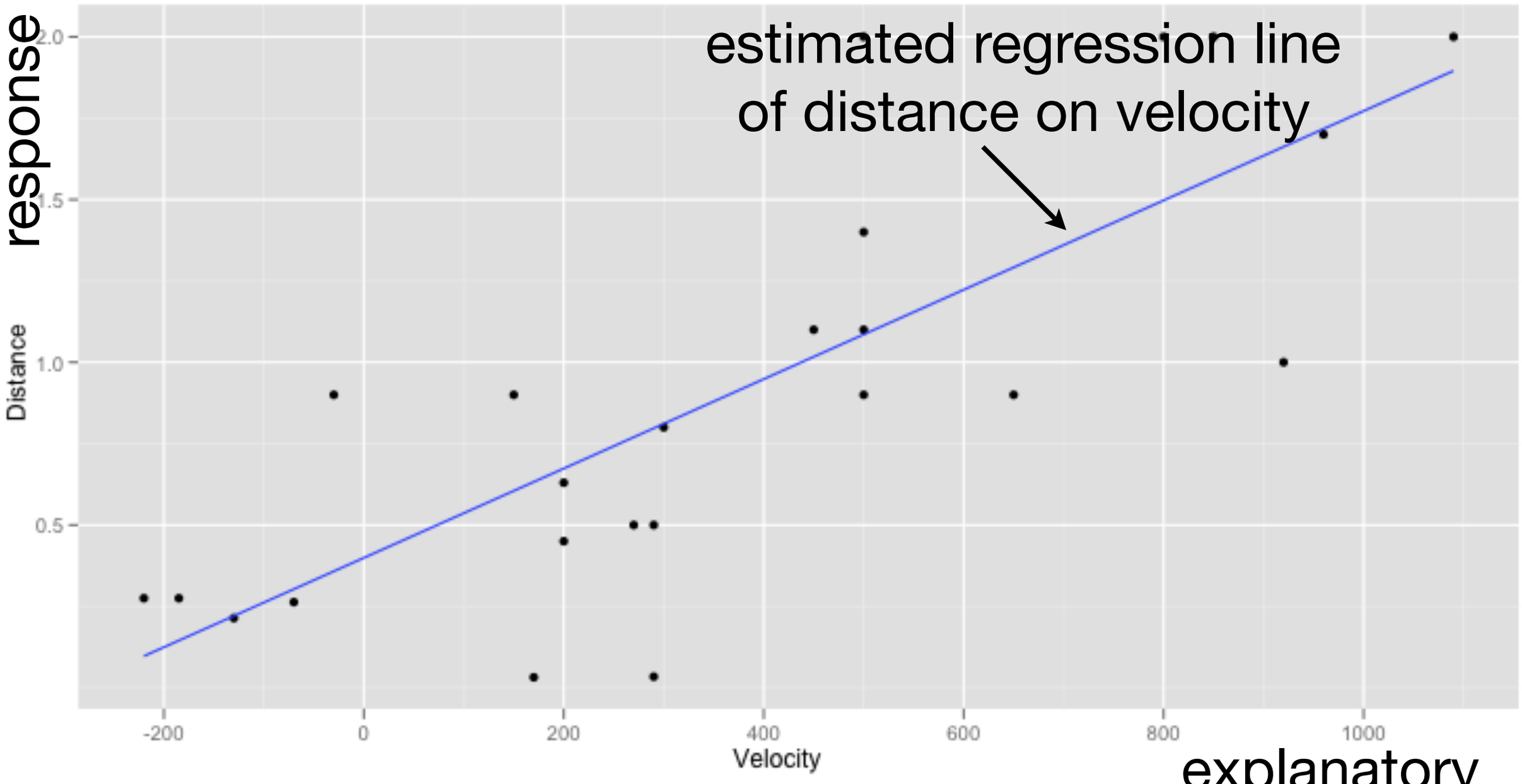
~~(the grouping variable)~~

many explanatory variables =
multiple linear regression -> ST512

We will be interested in predicting the **mean response** at each value of the **explanatory** variable with a **straight line**.

Big Bang (case0601)

```
qplot(Velocity, Distance, data = case0701) +
    geom_smooth(method = "lm", se = FALSE)
```



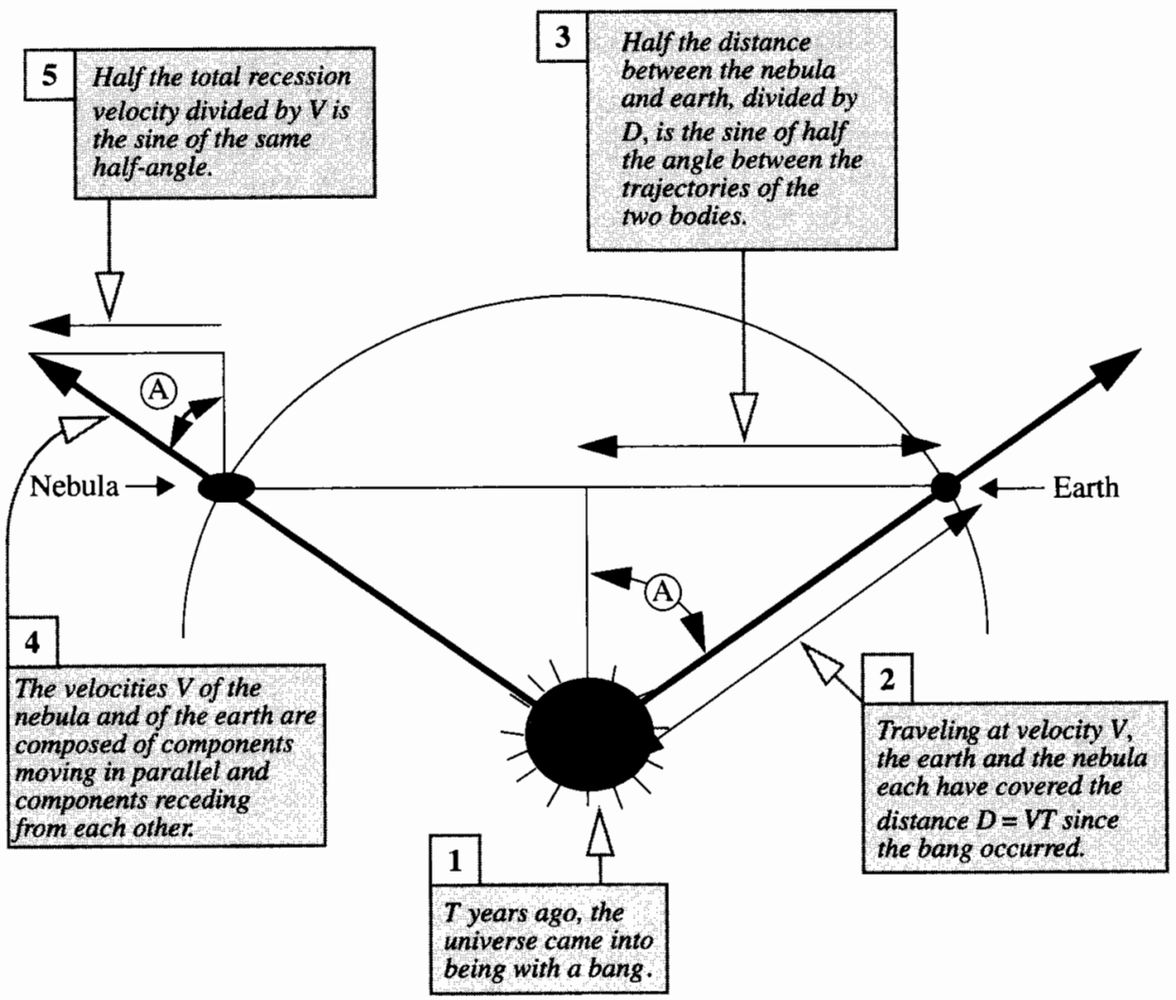
7.1.1 The Big Bang—An Observational Study

Edwin Hubble used the power of the Mount Wilson Observatory telescopes to measure features of nebulae outside the Milky Way. He was surprised to find a relationship between a nebula's distance from earth and the velocity with which it was going away from the earth. Hubble's initial data on 24 nebulae are shown as a *scatterplot* in Display 7.1. (Data from E. Hubble, "A Relation Between Distance and Radial Velocity Among Extra-galactic Nebulae," *Proceedings of the National Academy of Science* 15 (1929): 168–73.) The horizontal axis measures the recession velocity, in kilometers per second, which was determined with considerable accuracy by the red shift in the spectrum of light from a nebula. The vertical scale measures distance from the earth, in megaparsecs: 1 megaparsec is 1 million parsecs, and 1 parsec is about 30.9 trillion kilometers. Distances were measured by comparing mean luminosities of the nebulae to those of certain star types, a method that is not particularly accurate. The data are shown later in this chapter as Display 7.8.

Display 7.

The apparent statistical relationship between distance and velocity led scientists to consider how such a relationship could arise. It was proposed that the universe came into being with a Big Bang, a long time ago. The material in the universe traveled out from the point of the Big Bang, and scattered around the surface of an expanding sphere. If the material were traveling at a constant velocity (V) from the point of the bang, then the earth and any nebulae would appear as in Display 7.2.

Display 7.2 Big Bang theory model for distance-velocity relationship of nebulae



The distance (Y) between them and the velocity (X) at which they appear to be going away from each other satisfy the relationship

$$(Y/2)/VT = (X/2)/V = \sin(A),$$

where A is half the angle between them. In that case,

$$Y = TX$$

is a straight line relationship between distance and velocity. The points in Display 7.1 do not fall exactly on a straight line. It might be, however, that the *mean* of the distance measurements is TX. The slope parameter T in the equation $\text{Mean}\{Y\} = TX$ is the time elapsed since the Big Bang (that is, the age of the universe).

Several questions arise. Is the relationship between distance and velocity indeed a straight line? Is the y-intercept in the straight line equation zero, as the Big Bang theory predicts? How old is the universe?

Notation

We will use the symbol Y for the **response** variable.

We will use the symbol X for the **explanatory** variable.

When talking about observed data, Y_i and X_i are the observed response and explanatory variable for **i^{th} observation**.

i can be between 1 and n (the sample size)

$$\mu\{Y|X\}$$

is the mean of Y as a function of X.

is the mean of the response as a function of the explanatory variable.

$\mu\{Y|X\} = E\{Y|X\}$ if you've taken ST521

The simple linear regression model

Parameters

$$\mu\{Y|X\} = \beta_0 + \beta_1 X$$

Intercept Slope

The **mean** response is a **straight line** function of the explanatory variable. + some other assumptions

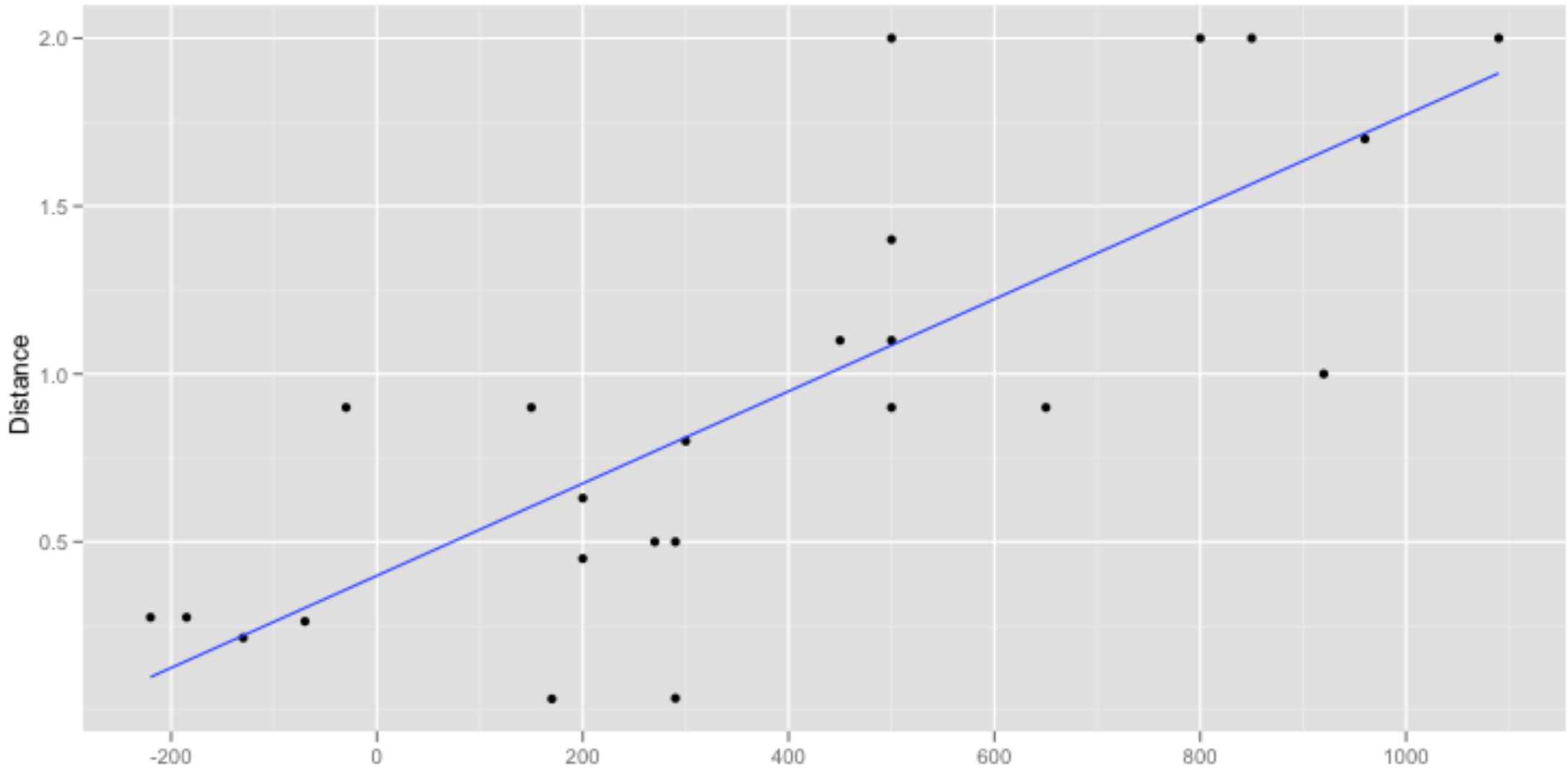
Describes the relationship between the response and explanatory variable with two parameters.

Chapter 7 is devoted to estimating β_0 and β_1 and their statistical properties

Intercept and Slope

The **intercept** gives the mean response at an explanatory value of zero, $\mu\{Y|\ X = 0\}$.

The **slope** gives the associated **change in the mean response** for a **1 unit increase** in the explanatory variable.



estimate of $\beta_0 = 0.399$ intercept

estimate of $\beta_1 = 0.0014$ slope

We estimate the **mean** distance of a nebula travelling at 0 km/sec to be 0.399 parsecs from Earth.

We estimate that an increase in velocity of 1 km/sec is associated with an increase in **mean** distance from Earth by 0.0014 parsecs.

Wednesday(?) beers

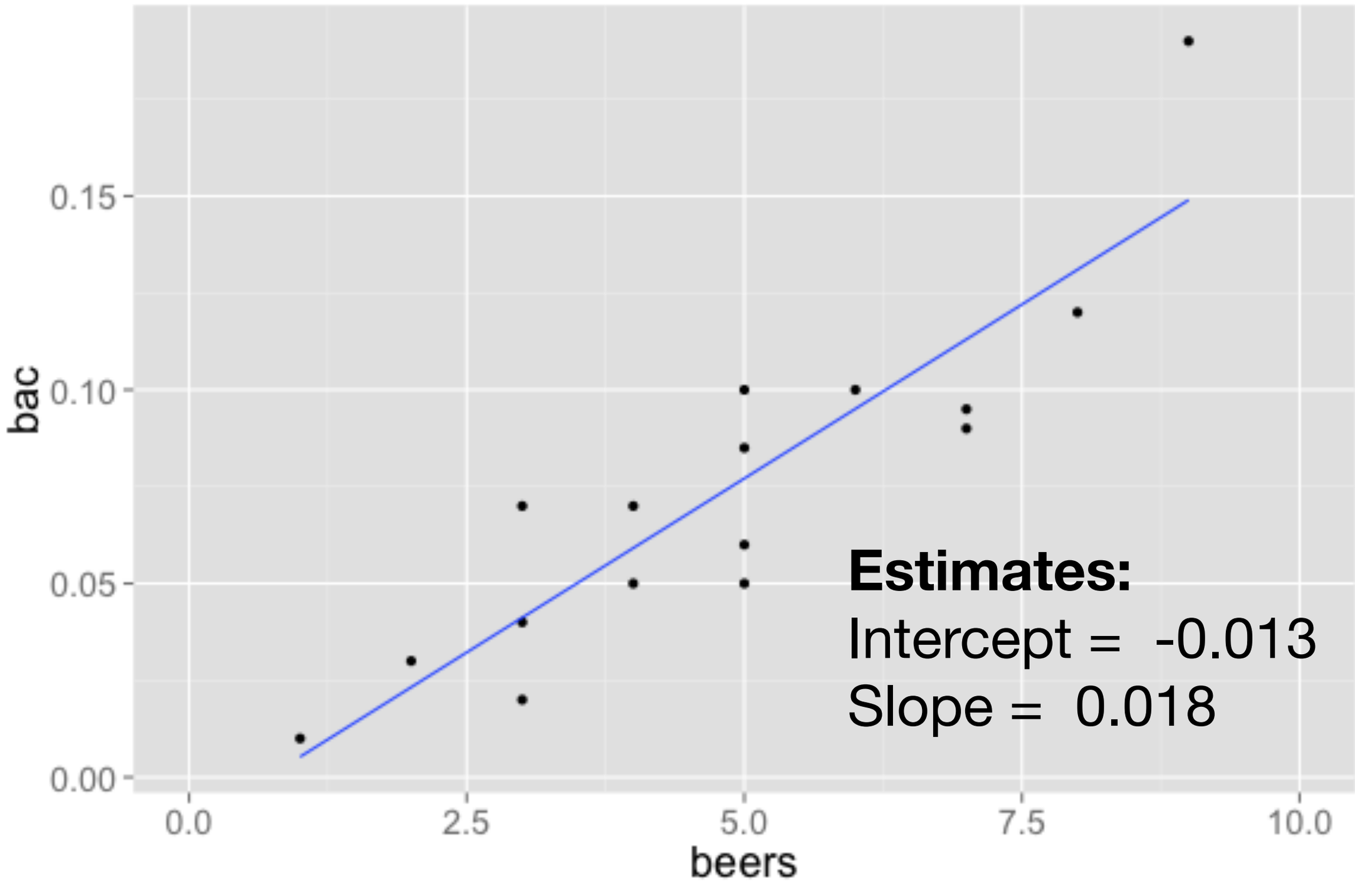
A study conducted at Ohio State University randomly assigned student volunteers to a number of cans of beer to drink.

Thirty minutes later, a police officer measured their blood alcohol content, BAC.

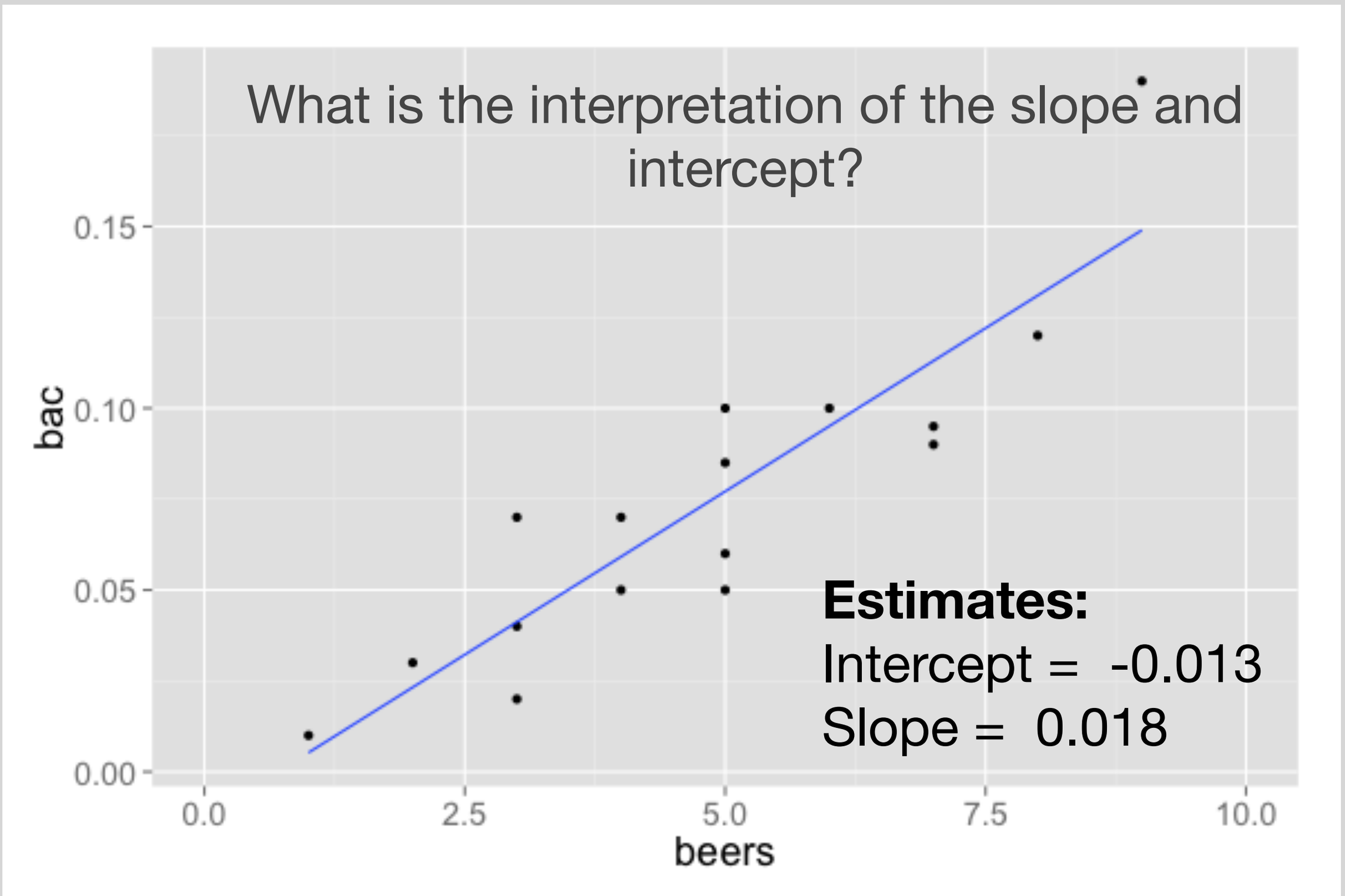
The officers would like to investigate the relationship between blood alcohol content and number of beers.

Beer and BAC

```
qplot(beers, bac, data = beer) +
    geom_smooth(method = "lm", se = FALSE)
```



Your turn


What is the interpretation of the slope and intercept?
0.15
0.10
0.05
0.00
bac
0.0
2.5
5.0
7.5
10.0
beers
Estimates:
Intercept = -0.013
Slope = 0.018

Your turn

Intercept: -0.013

We estimate, that **mean** BAC after drinking 0 beers is -0.013 %. (correct interpretation of model).

Slope: 0.018

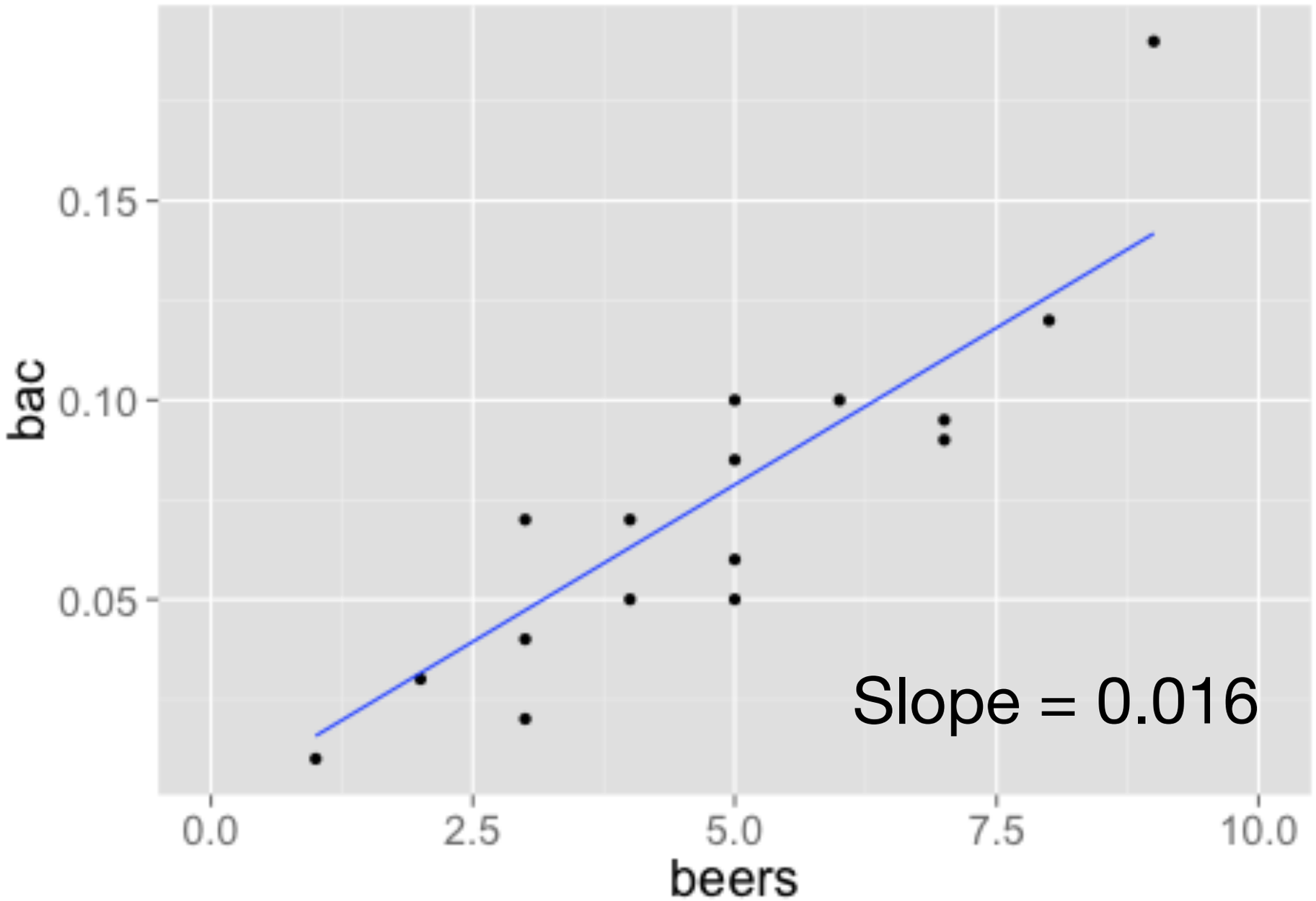
We estimate that the mean BAC will increase 0.018 percentage points with each additional beer that is consumed. (correct for this study).

We estimate that with each additional beer that is consumed there is an associated increase in the mean BAC of 0.018 percentage points. (observational study wording).

not examinable

Force intercept = 0

```
qplot(beers, bac, data = beer) +
    geom_smooth(method = "lm", se = FALSE,
                formula = y ~ x - 1 )
```



When **X = 0,** falls outside the explanatory values of interest, it doesn't make sense to interpret the intercept.

e.g. height versus weight, how much on average does someone weigh who is 0cm tall?

You can change the scale on the slope if it's easier to understand,

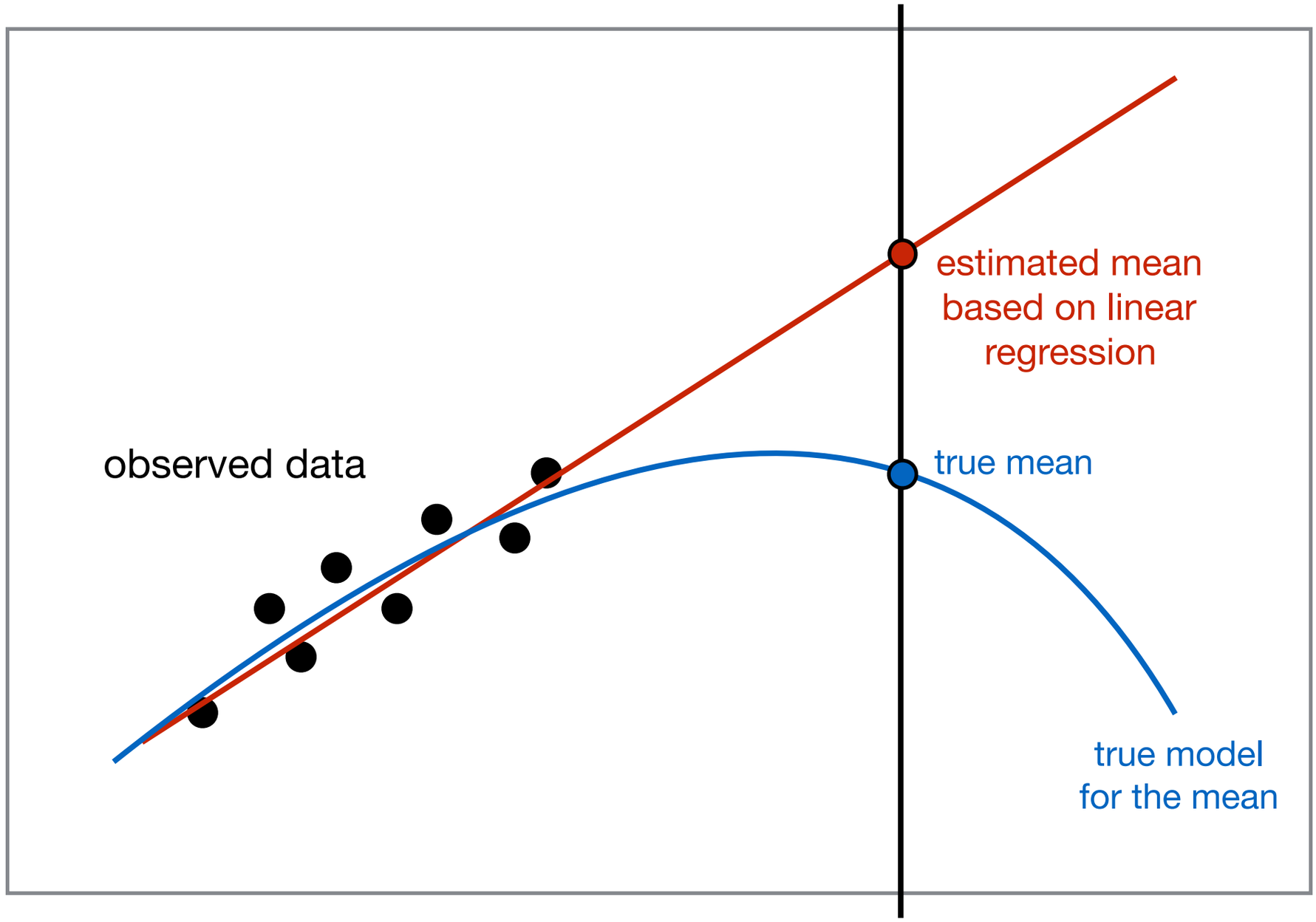
a change in the explanatory of Z units results in a change in the mean response of (Z × slope) units.

Interpolation and extrapolation

Using the straight line model we can predict the response for any value of the explanatory variable.

Interpolation is when we predict the response using a value **within** the range of measured explanatory values.

Extrapolation is when we predict the response using a value **outside** the range of measured explanatory values.

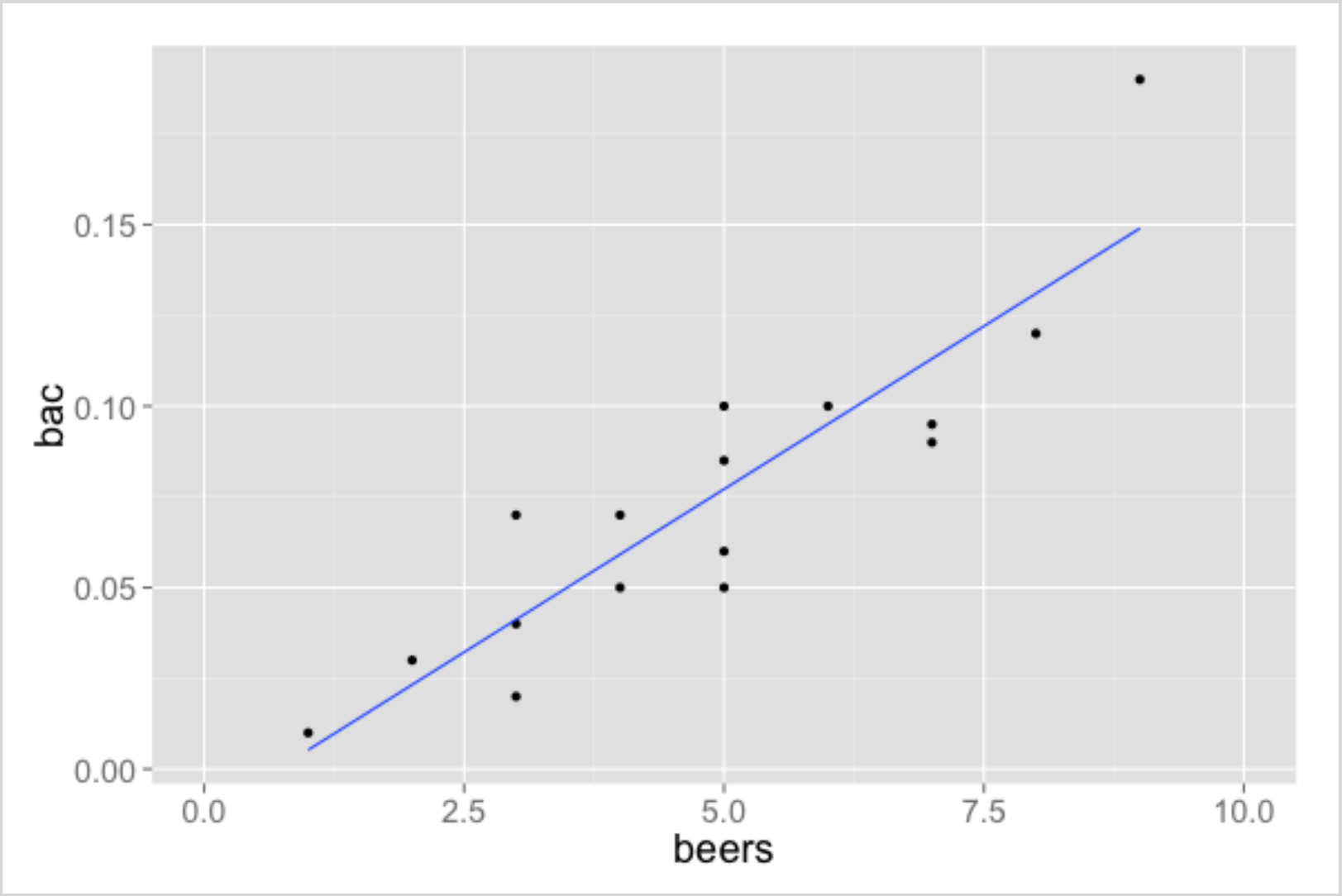


Extrapolation is dangerous because we can’t check the adequacy of our model for the values of interest using the data at hand.

Your turn

Interpolation or extrapolation?

Mean BAC after 3.5 beers
Mean BAC after 10 beers
Mean BAC 0.5 beers



Next...

How do we estimate the slope and intercept?

What kind of uncertainty do we have about those estimates?

How do we make inference about the **mean response**?

How do we make inference about the **response**?

$\sum$

A bit of summation notation next week. You should be comfortable with expressions like:

$$\sum_i Y_i = Y_1 + Y_2 + \ldots + Y_n$$

$$1/n \sum_i Y_i = \bar{Y}$$

$$\sum_i (Y_i - \bar{Y}) = (Y_1 - \bar{Y}) + (Y_2 - \bar{Y}) + \ldots + (Y_n - \bar{Y})$$

$$\sum_i (X_i - \bar{X})(Y_i - \bar{Y})$$

$$= (X_1 - \bar{X})(Y_1 - \bar{Y}) + \ldots + (X_n - \bar{X})(Y_n - \bar{Y})$$

You won't have to calculate anything (R will do the work) but you need to be comfortable seeing them.