Stat 412/512

Review of Simple Linear Regression

Jan 7 2015

Charlotte Wickham

stat512.cwick.co.nz

Announcements

TA's
Katie 2pm lab

Ben 5pm lab

Joe noon & 1pm lab

TA office hours Kidder M111
Katie Tues 2-3pm

Ben Thur noon-1pm

Joe Mon 3-4pm & Thur 9-10am

Reminder my **ST512** office hours :

Tuesday 3-5pm Cordley 3003

Friday 11-noon Kidder 76

First homework posted!

Review

Simple linear regression:

model for the mean

interpreting intercept and slope

assumptions and residuals

R output

Types of statistical inference

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

Y

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

X

The **simple linear regression** model

Parameters

$$\mu\{Y|X\} = \underset{\text{Intercept}}{\beta_0} + \underset{\text{Slope}}{\beta_1} X$$

The mean response as a function of the explanatory variable is a straight line.

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

Intercept and Slope

The **intercept** gives the mean response at an explanatory value of zero.

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

Your turn

A simple linear regression of stopping distance (ft) on speed (mph) :

$\mu\{\text{dist} \mid \text{speed}\} = \beta_0 + \beta_1\text{Speed}$

from 50 cars, gave the following estimates:

Parameter	Estimate	95% confidence interval
Intercept, β_0	-17.6	-31.2, -4.0
Slope, β_1	3.9	3.1, 4.8

On your own, write down a two sentence summary interpreting the slope estimate.

It is estimated that

① For every 1 mph increase in speed the mean stopping distance increases by 3.9 feet.

The mean stopping distance increase by 3.9 feet, ...

② With 95% confidence, . . .

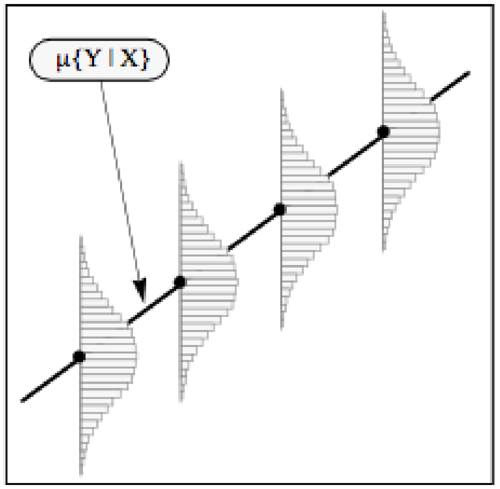
increases between 3.1 & 4.8 feet.

Your turn

It is estimated that for every one mile an hour increase in speed the **mean** stopping distance increases by 3.9 feet (95% CI 3.1 to 4.8).

With 95% confidence, for every one mile an hour increase in speed the **mean** stopping distance increases by between 3.1 and 4.8 feet

Assumptions



1. The means of the subpopulations fall on a **straight line** function of the explanatory variable. $\mu\{Y \mid X\} = \beta_0 + \beta_1 X$

2. The subpopulations have the **same standard deviation**, σ. $\mathrm{Var}\{Y \mid X\} = \sigma^2$

3. At each value of the explanatory, the response has a **Normal** distribution.

4. Observations are **independent**.

There are no assumptions on the explanatory variable!

An alternative view of
the assumptions

Assumptions

If the linear effect of the explanatory variable is removed from the response, then what's left should be:

1. **Normally distributed**
2. Have **mean zero**.
3. Have **standard deviation**, σ.
4. Be **independent**.

I.e. (Y - (β_0 + β_1 X)) are independent Normal with mean zero and standard deviation σ

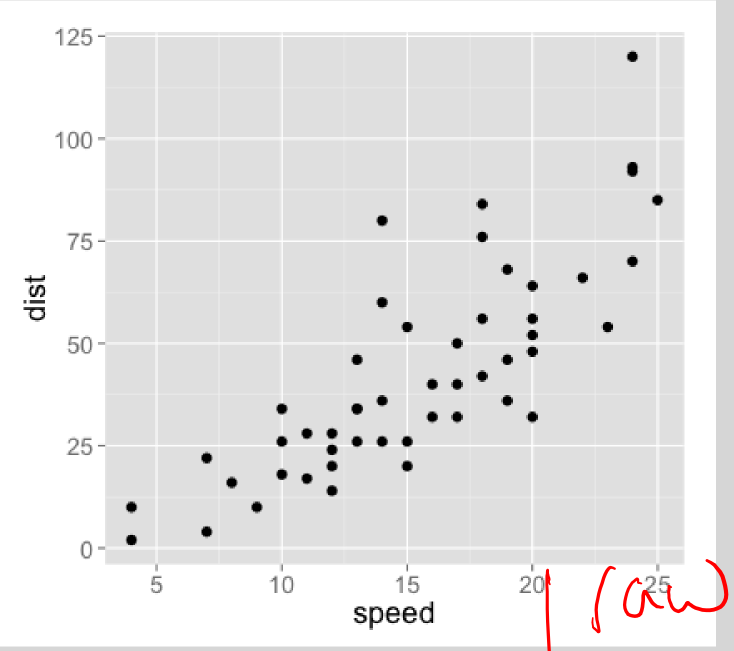
Your turn

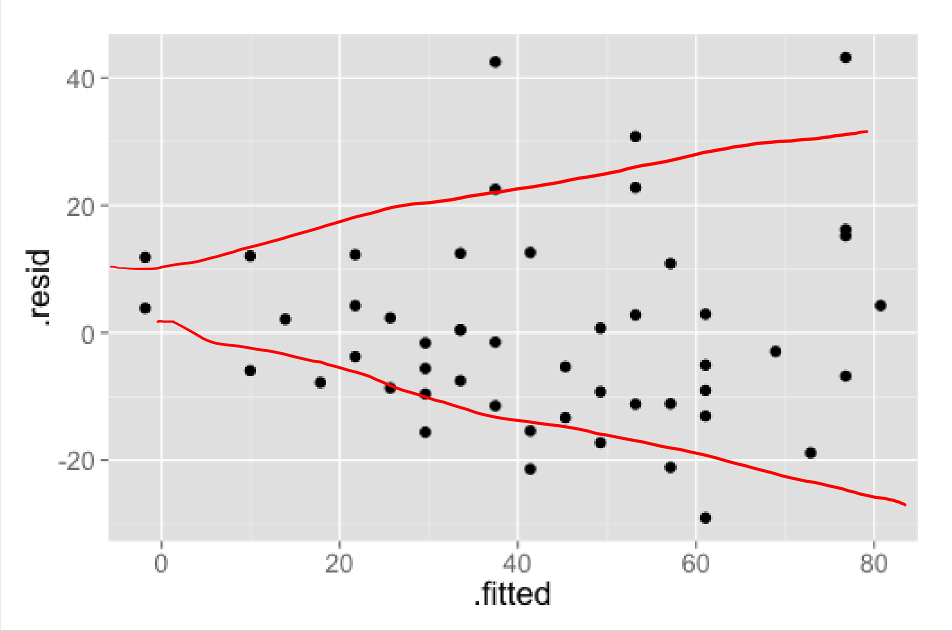
Someone suggests that we should have done the regression of log(dist) against log(speed) (because the relationship probably isn’t additive, e.g. the difference in stopping distance is probably bigger between 60 and 70 mph than it is between 10 and 20 mph)

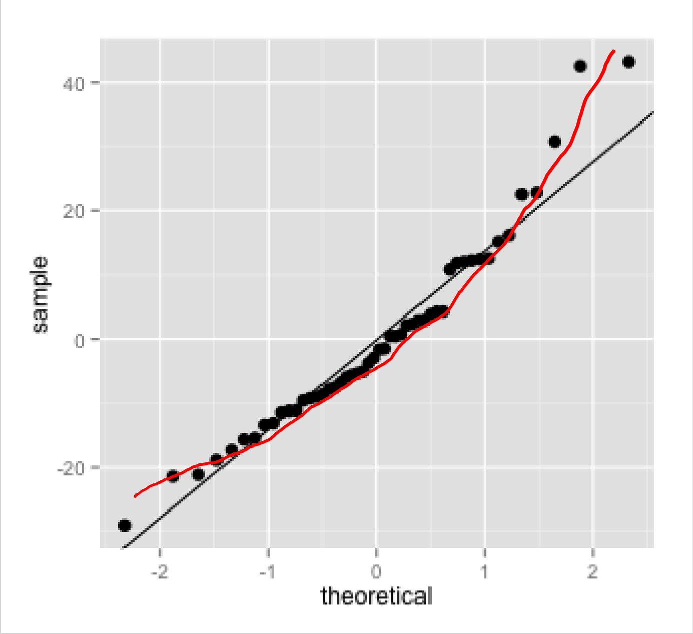
How would you decide whether their suggestion is a good one?

Your turn

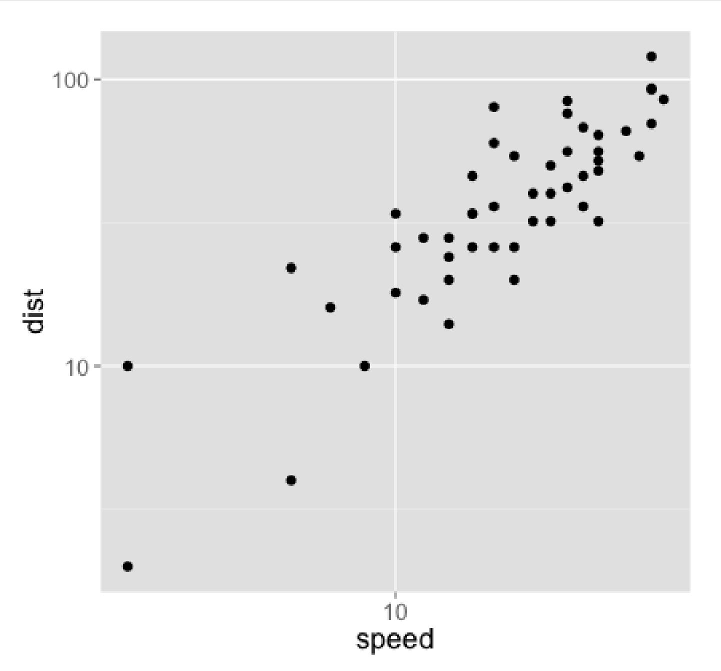
dist ~ speed

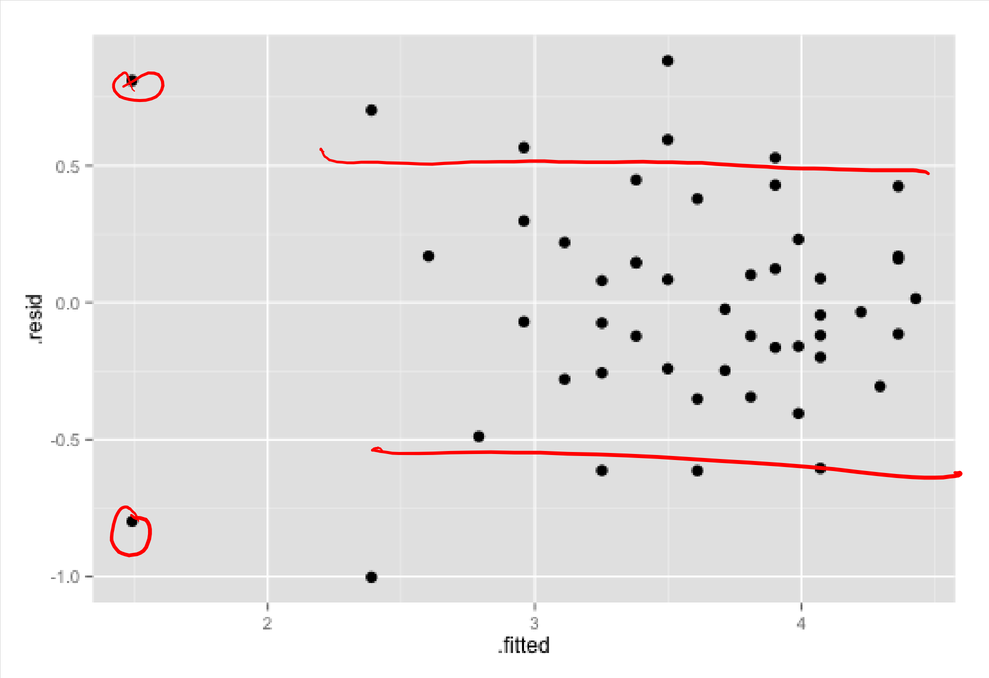


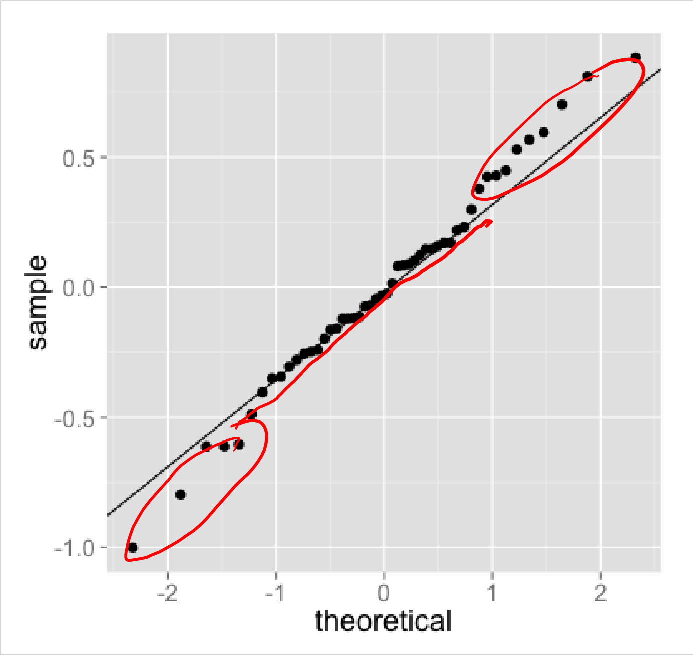




log(dist) ~ log(speed)







Inference in simple linear regression

by least squares
↓

t-ratios of the estimates of the slope, intercept, mean response, and predicted response all have a **Student's t-distribution** with **n - 2** degrees of freedom.

Competing (nested) models can be compared using an extra Sum of Squares F-test.

Your turn

```
> summary(fit2)

Call:
lm(formula = log(dist) ~ log(speed), data = cars)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.00215 -0.24578 -0.02898  0.20717  0.88289 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.7297     0.3758  -1.941   0.0581 .  
log(speed)    1.6024     0.1395  11.484 2.26e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4053 on 48 degrees of freedom
Multiple R-squared:  0.7331,	Adjusted R-squared:  0.7276 
F-statistic: 131.9 on 1 and 48 DF,  p-value: 2.259e-15
```

1. What test is this p-value for?

Null: slope = 0

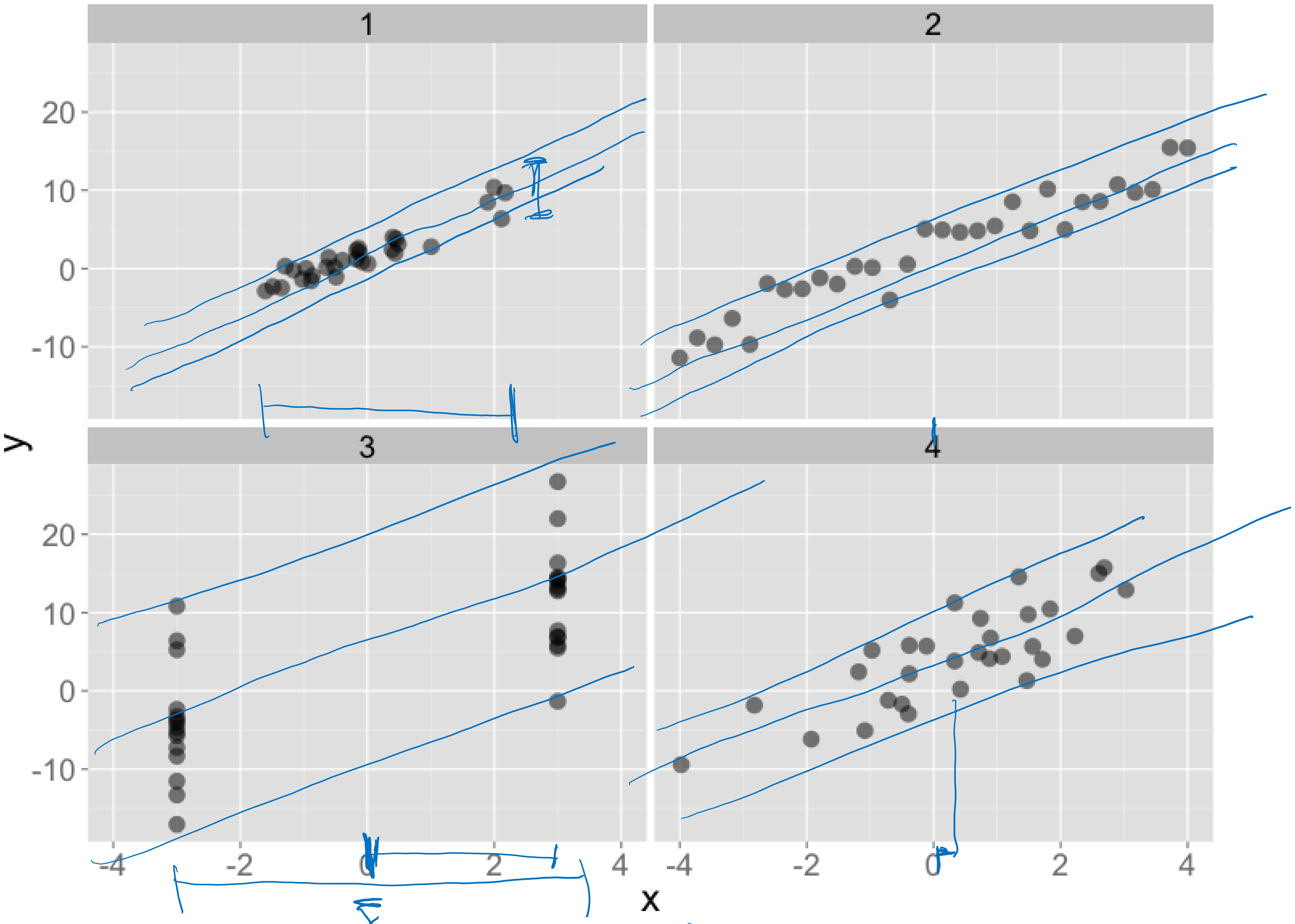
2. How is this t-statistic calculated?

$$\frac{\text{Estimate}}{\text{Std. Error}}$$

3. What test is this p-value for?

Regression ANOVA

Compares regression to equal means



Which dataset has the largest $\hat{\sigma}$ (subpopulation response sd)?

Which dataset has the largest s_X (explanatory sd)?

A larger sample (larger n) **does not decrease σ**, we simply get a more precise estimate.

A larger sample (larger n) **does decrease the standard error** on the slope and intercept estimates.

σ is often outside the control of the researcher, but sometimes it can be decreased by improving the measurement of the response.

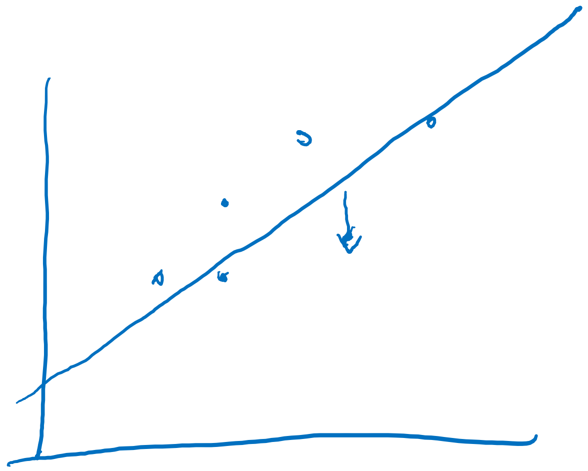
Eg. using a digital thermometer, rather than one of those color change magnets to measure temperature.

The standard error of the slope and intercept also depend on how you choose X (if you can). Normally it's a balance between low standard error, and the ability to check assumptions and will depend on your questions of interest.

The **fitted value** describes the **estimated mean** for an observation. For the ith observation the fitted value is,

$$\text{fitted}_i = \hat{\mu}\{Y_i|X_i\} = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

The **residual** is the difference between the observed response and it's fitted value



$$\text{residual}_i = Y_i - \text{fitted}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

The appropriateness of the linear regression model is checked by looking at plots of the residuals

Review in lab and homework

Population and Causal Inference

Experimental unit: the object to which the treatment is applied.

Sampling unit: the object that is selected from a population.

Observational unit: the object that measurements are taken on.

Population inference (statistical inference beyond the units in the study to a wider population) is justified if the sampling units in the study are a **random sample from the population**.

Causal inference (statistical inference that the treatment caused the differences observed in the study) is justified if the experimental units in the study were **randomly assigned to treatments**.

In simple (and multiple) linear regression

In regression, the “treatments” are specific values of the explanatory variables. Experimental units must be randomly assigned to the values of the explanatory variables for causal inference to be valid.

To infer a cause-effect relationship the researcher must decide the levels of the explanatory variable they are going to measure, and then randomly assign their subjects to those levels.

If the subjects are not randomly assigned to predetermined levels, or the explanatory variables are not under the control of the researcher (i.e. they simply observe it) **causal inference is not justified**.

We can still talk about a relationship or association between the variables but we must avoid causal language, (e.g. avoid saying “X increases Y”)

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

George E. P. Box

Regression models, like all models, are never going to be a perfect description of reality.

We will always keep two things in mind:

* we want our models to be **good** approximate descriptions of reality (i.e. not too wrong)

* we want our models to be **useful** in answering our scientific questions of interest

General approach

for hypothesis testing and estimation

1. Fit a regression model that is designed to answer your questions.

2. Check the validity of the model

If it's OK, proceed. If it isn't OK, refine the model.

3. Answer your questions of interest using the model and make appropriate inferences.