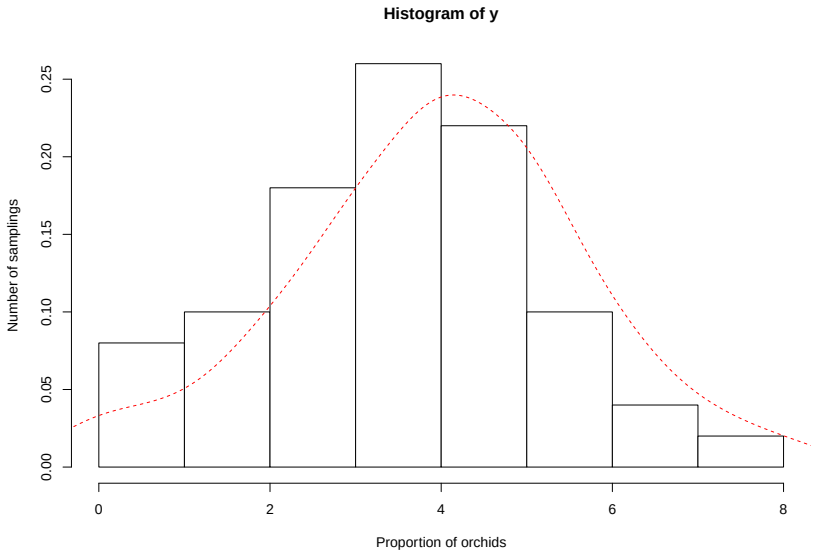
Introduction to Linear Models

Bert van der Veen

Department of Mathematical Sciences, NTNU

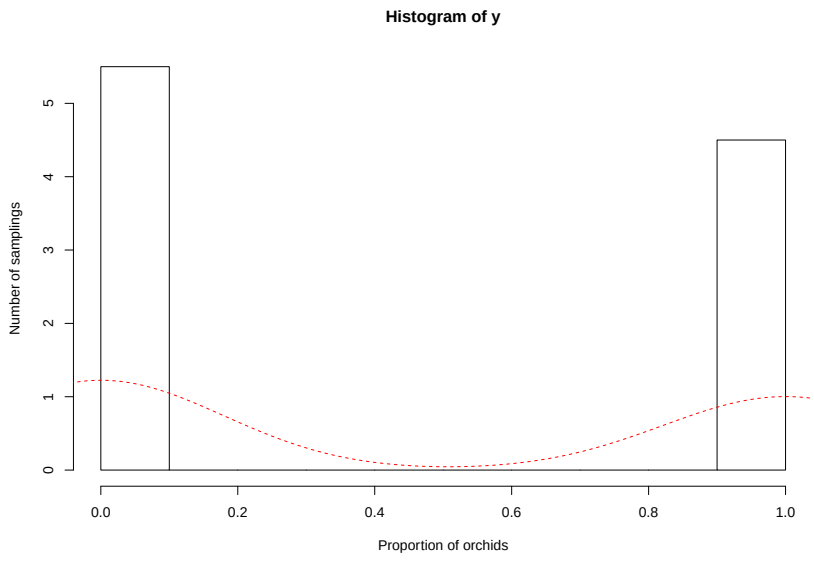
The orchids example

```
set.seed(12345) # For reproducibility
n.times <- 50;n.picks=10;p.orchid <- 0.4
y <- rbinom(n.times, size = n.picks, prob = p.orchid)
```



The orchids example

```
set.seed(12345) # For reproducibility
n.times <- 100;n.picks=1;p.orchid <- 0.4
y <- rbinom(n.times, size = n.picks, prob = p.orchid)
```

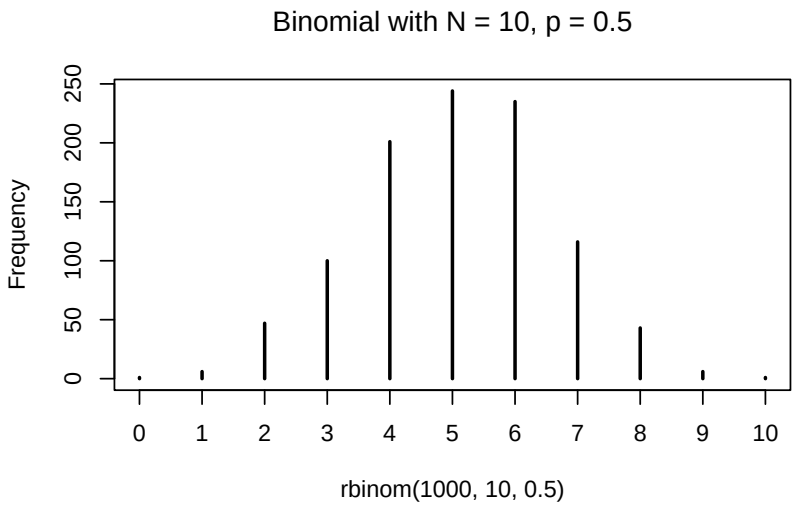


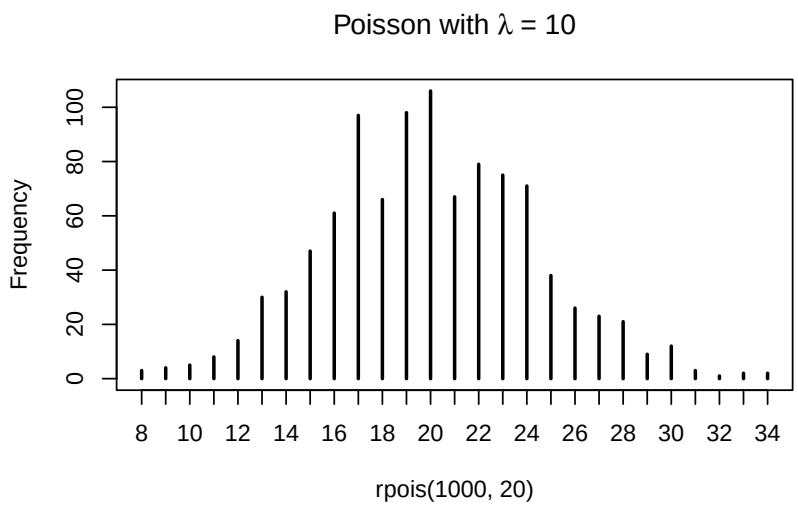
The orchids example

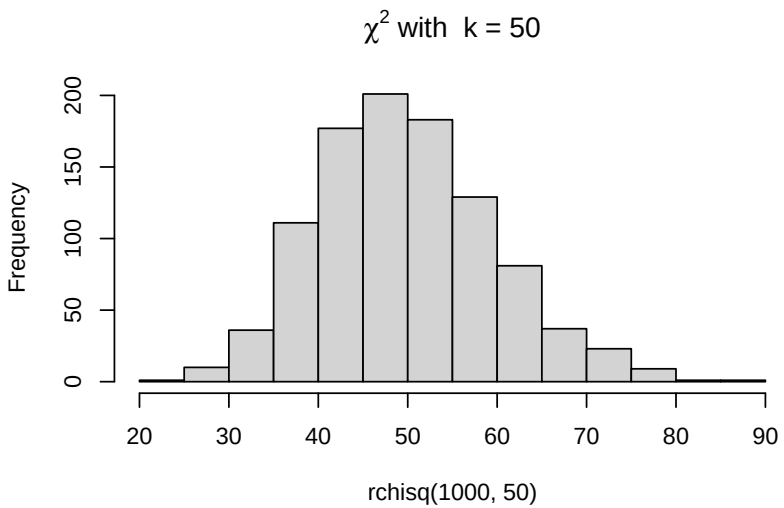
The Binomial distribution is approximately normal for large n_{picks} and π away from 0 and 1

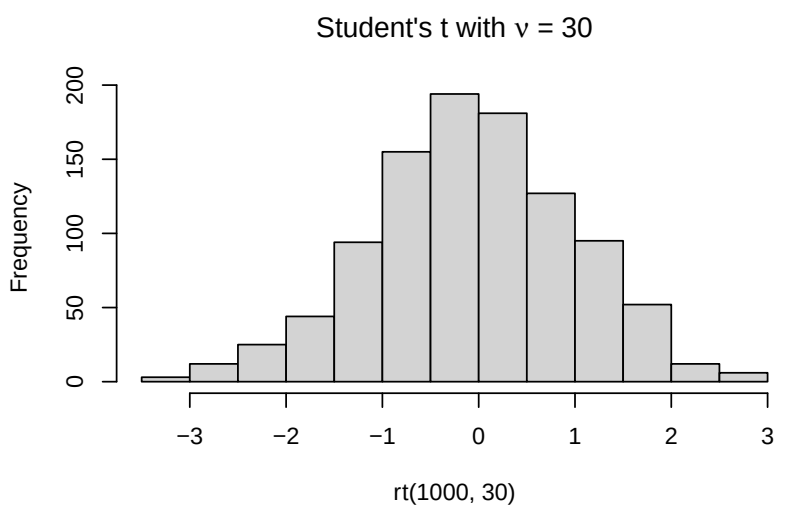
Many distributions are approximately normal under some conditions

Normal approximation



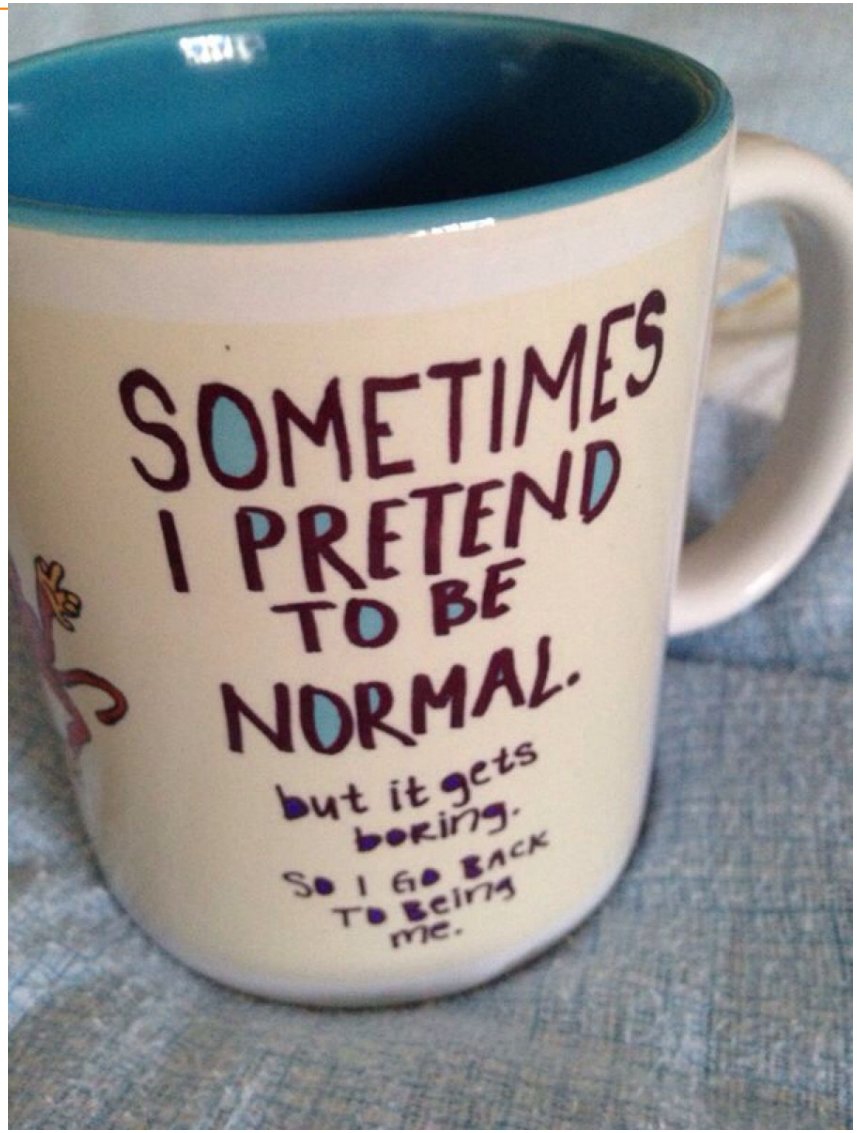






Normality

Real data is rarely normally distributed.
But it is a nice starting point when learning GLMs.
We can use it (e.g.,) when the mean is far enough from zero.



The normal distribution

$$f(y_i; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{(y_i - \mu)^2}{2\sigma^2} \right\} \tag{1}$$

The normal distribution (2)

Likelihood:

$$f(y_i; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{(y_i - \mu)^2}{2\sigma^2} \right\} \tag{2}$$

log-Likelihood:

$$\log\{f(y_i; \mu, \sigma)\} = -\frac{1}{2} \log(\sigma^2 2\pi) - \frac{(y_i - \mu)^2}{2\sigma^2} \tag{3}$$

Two parameters: μ and σ

- μ is the mean; the middle of the distribution
- σ is the standard deviation; it controls the width

Not only used for data, also the basis of many statistics (e.g., asymptotic sampling distributions)

Estimating μ

Same process as before: calculate gradient and find estimator

$$\frac{\partial \log\{\mathcal{L}(\mathbf{y}; \hat{\mu}, \sigma)\}}{\partial \mu} = \frac{1}{2\sigma^2}\left(2\sum_{i=1}^{n} y_i - 2n\mu\right) \tag{4}$$

Giving..

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{5}$$

It is a linear function of y_i so it follows a normal distribution.

Estimating σ^2

Same process as before. MLE is biased so gets a small correction.

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{\mu})^2 \tag{6}$$

Is it a quadratic function of y_i so is χ^2-distributed.

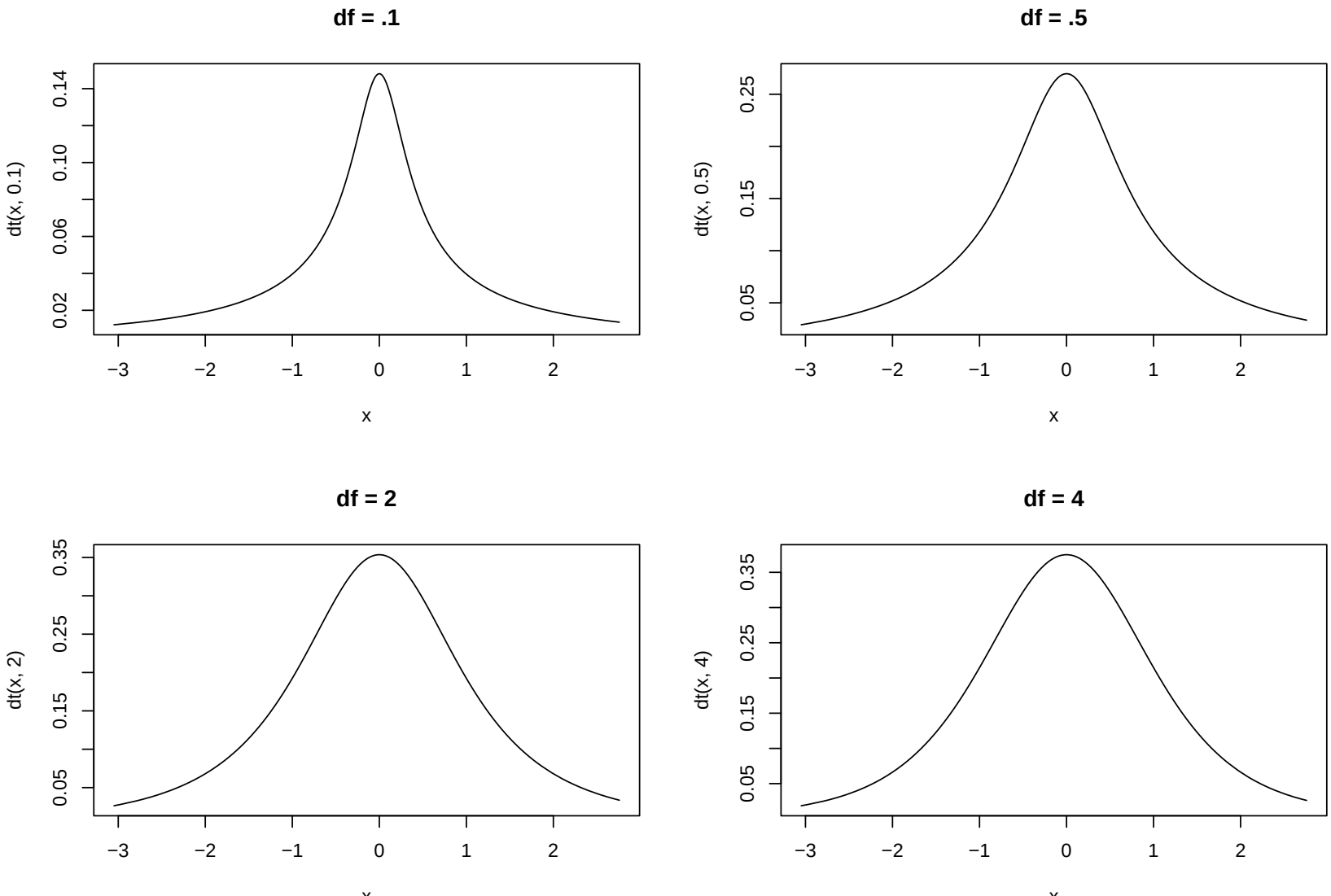
Uncertainty of $\hat{\mu}$

$$\begin{aligned} \text{var}(\hat{\mu}) &= \mathbb{E}(\hat{\mu}^2) - \mathbb{E}(\hat{\mu})\mathbb{E}(\hat{\mu}) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}(y_i^2) - \mathbb{E}(\hat{\mu})\mathbb{E}(\hat{\mu}) \\ &= \frac{1}{n}\sigma^2 \end{aligned} \tag{7}$$

- Depends on n (small n, large uncertainty)
- Depends on σ^2
- Which is estimated by $\hat{\sigma}^2$
- But that estimate also has uncertainty

So, we use the t-distribution to represent that additional uncertainty.

The t-distribution



The t-test

A t-test is used to test the mean difference of two groups

```
set.seed(12345)
y <- rnorm(10)
x <- rnorm(10, mean = 2)
t.test(x, y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 6.5336, df = 17.979, p-value = 3.872e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.641042 3.196802
## sample estimates:
##  mean of x  mean of y
##  2.2859778 -0.1329441
```

Example 1: Alaskan pipeline

Depth of pipeline defects

Example 1: the data

Data from http://www.itl.nist.gov/div898/handbook/

- 107 defects
- Measured in the field (ultrasonic)
- And in the lab re-measured

field.defect	lab.defect	batch
18	20.2	1
38	56.0	1
15	12.5	1
20	21.2	1
18	15.5	1
36	39.0	1
20	21.0	1

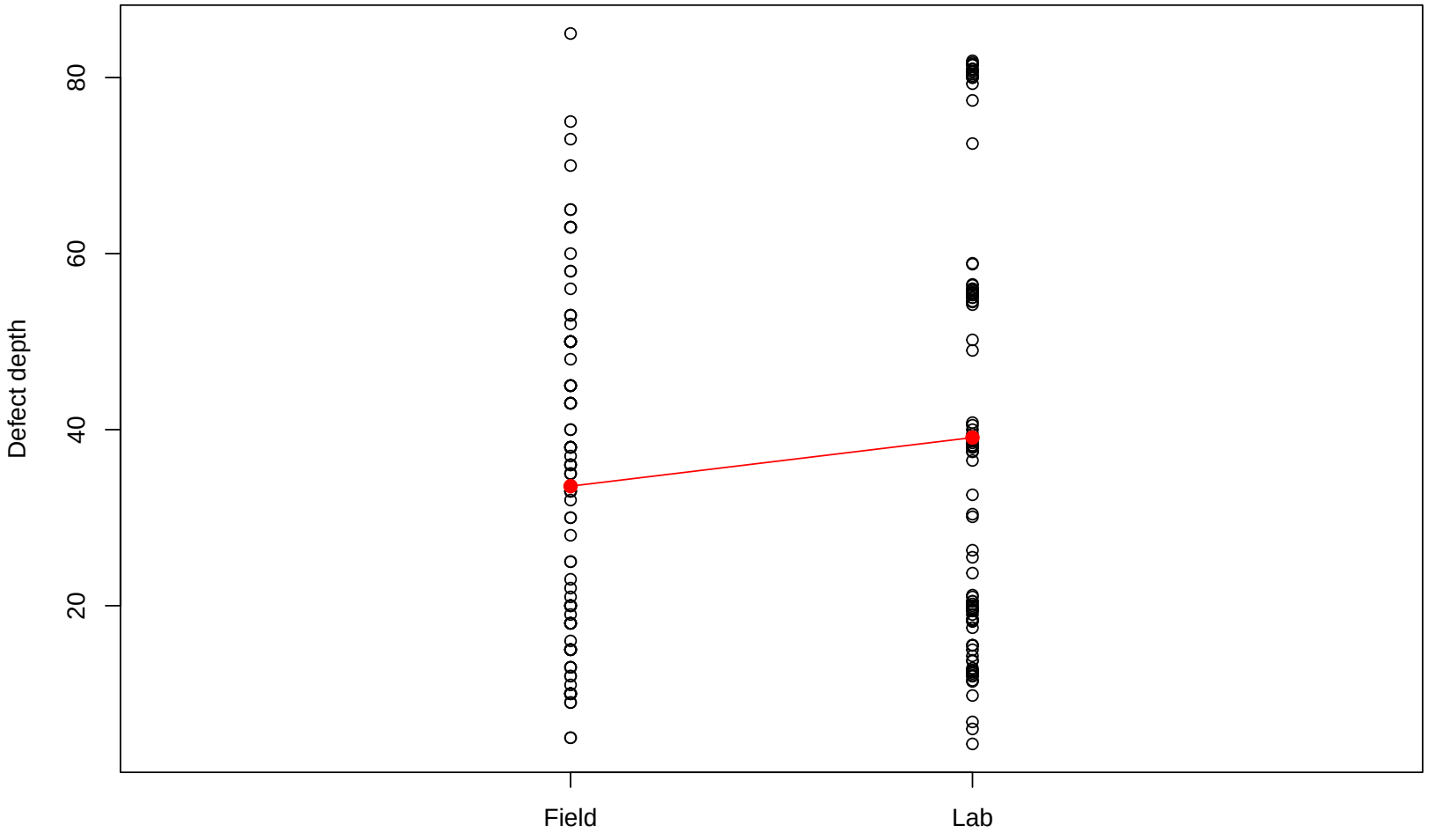
Alaska pipline: t-test

```
t.test(alaska.pipeline$field.defect,
       alaska.pipeline$lab.defect, paired=TRUE)
```

Alaska pipline: t-test output

```
##
##  Paired t-test
##
## data:  alaska.pipeline$field.defect and alaska.pipeline$lab.defe
## t = -6.4454, df = 106, p-value = 3.486e-09
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -7.217460 -3.821792
## sample estimates:
## mean difference
##       -5.519626
```

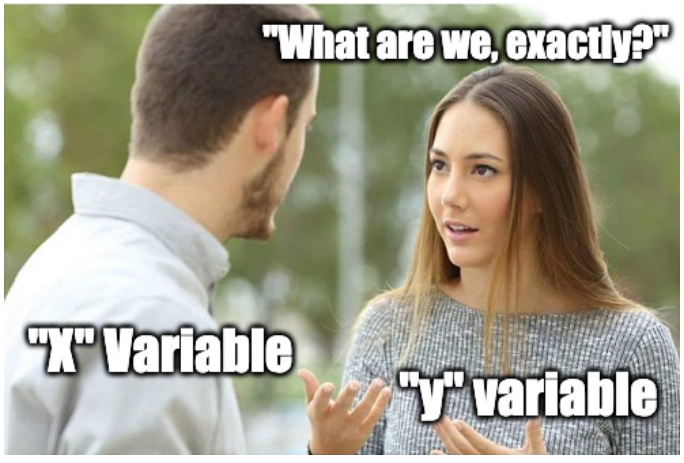
Alaska pipeline: t-test visually



What is a linear regression?

Models with a continuous **response variable** as a function of one or more **explanatory variable**. Variables are connected by linear equations.

- y_i: the **response variable**, can only be numerical
- x_i: the **explanatory variable**, can be categorical (0,1) or numerical



$$y_i = \alpha + x_i\beta + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{8}$$

Synonyms

- Covariate
- Predictor (variable)
- Explanatory variable
- Independent variable

They all refer to x_i.

What is the goal of regression?

We measure data y_i and want to infer its with x_i

Steps:

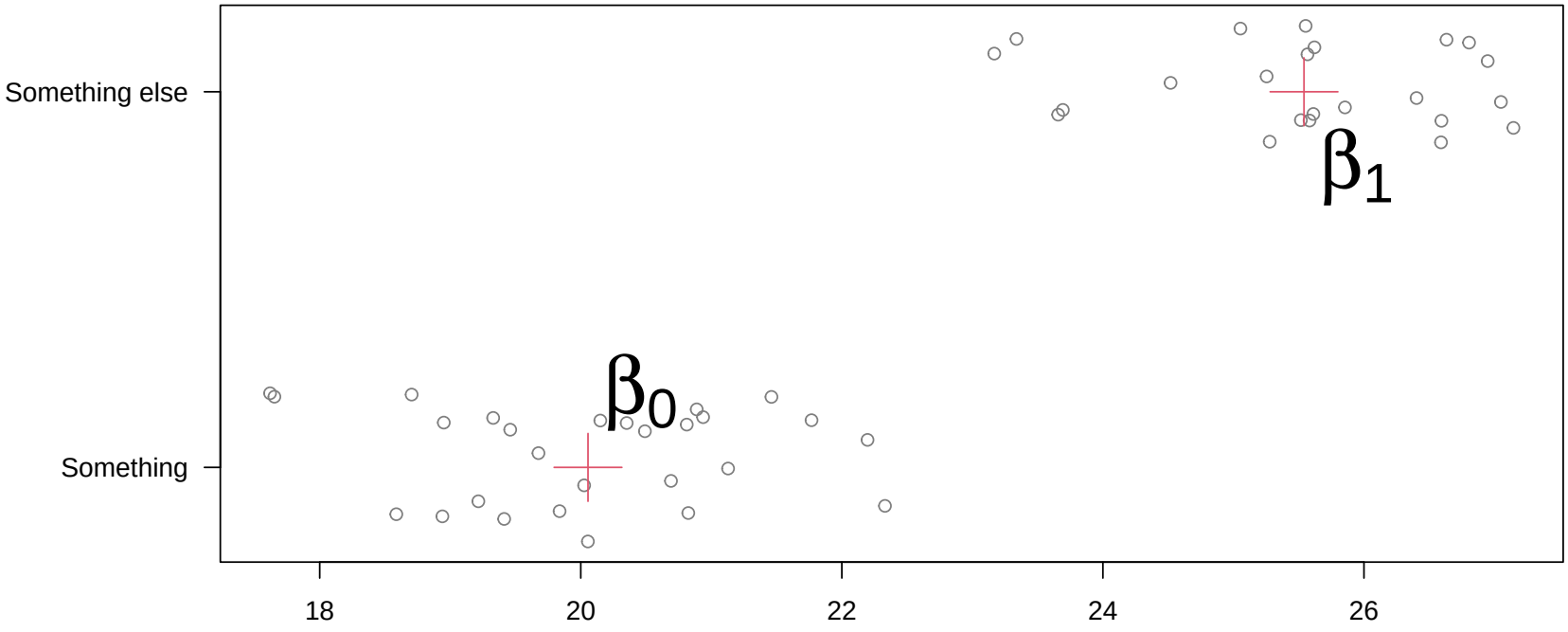
1) We decide on a model
2) We estimate the parameters
3) We check if it is a valid and good model
4) We draw our conclusion (with uncertainty)

Examples of linear models: categorical x_i

$$\mu_i = \begin{cases} \beta_0 & \text{if } X_i = 0 \\ \beta_1 & \text{if } X_i = 1 \end{cases} \qquad y_i = (1 - x_i)\beta_0 + x_i\beta_1 + \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

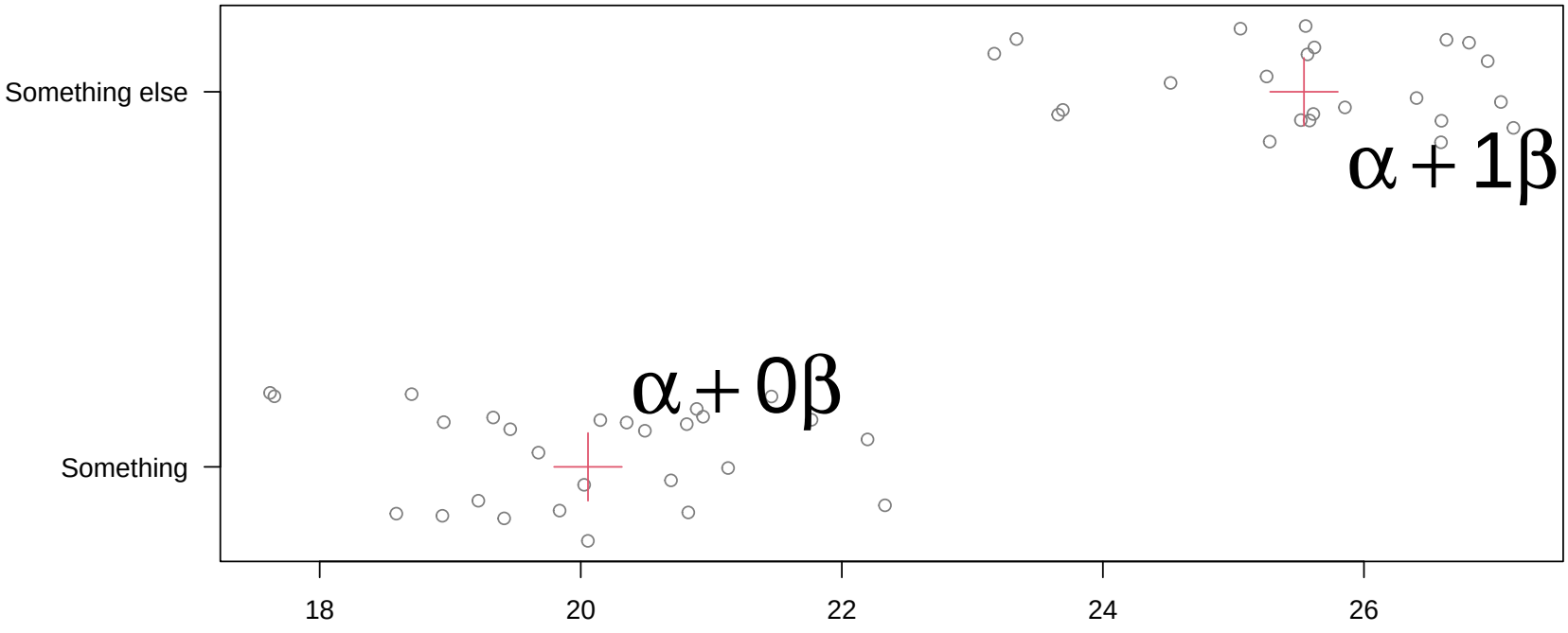
$$\mu_i = \begin{cases} \alpha & \text{if } X_i = 0 \\ \alpha + \beta & \text{if } X_i = 1 \end{cases} \qquad y_i = \alpha + x_i\beta_1 + \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Examples of linear models: categorical x_i



- β_0 is the group 1 mean
- β_1 is the group 2 mean

Examples of linear models: categorical x_i



- α is the mean of the first group
- β is the deviation from the mean of the first group

Linear regression and the t-test

Sounds familiar to the t-test?

Linear regression and the t-test

Sounds familiar to the t-test?

- t-test is a type of linear regression
- Namely where we have a categorical covariate with two categories

Alaska pipline: t-test as linear regression

$$depth_i = \begin{cases} \alpha & \text{if } measured\ in\ the\ field \\ \alpha + \beta & \text{if } measured\ in\ the\ lab \end{cases} + left.over.stuff_i$$

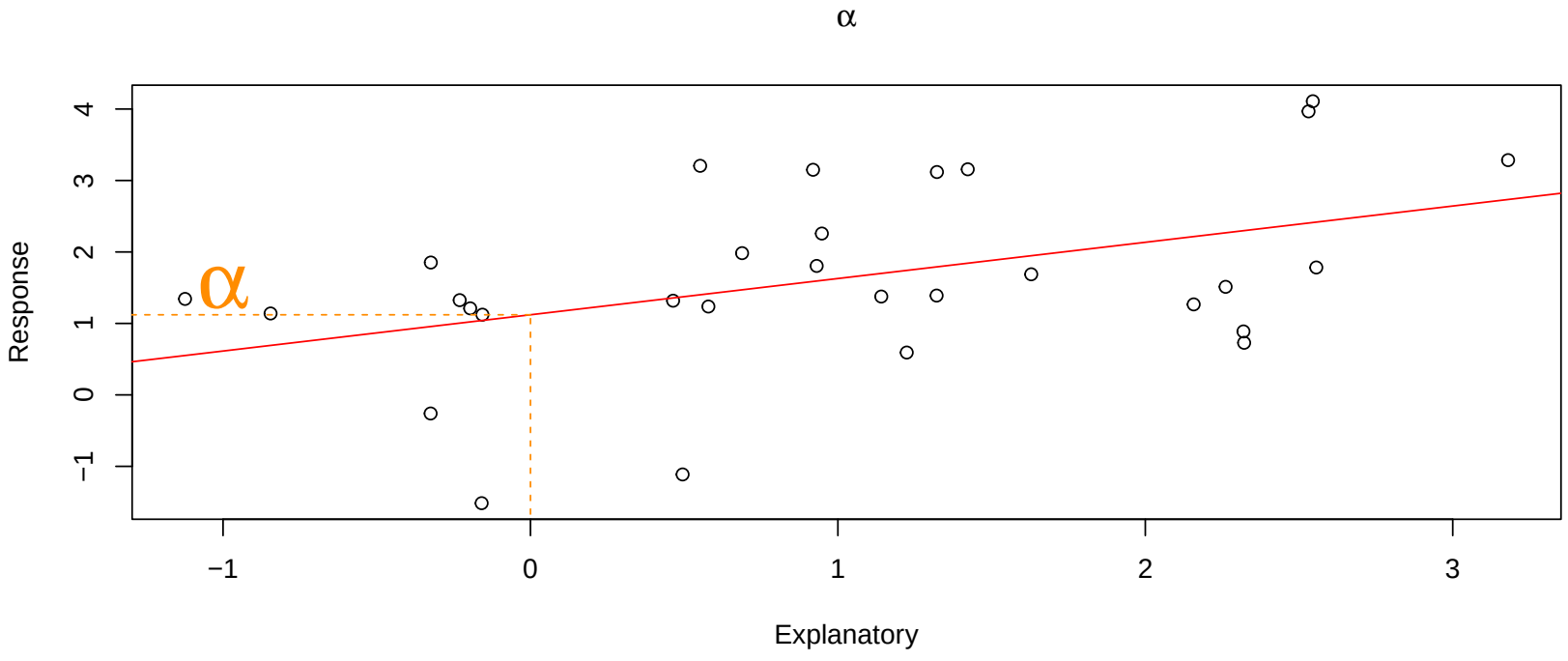
So, β is how much deeper (-) or less deep (+) the defect is detected when measured in the lab

Alaska pipline: t-test with `lm()`

```
alaska <- rbind(data.frame(id = 1:nrow(alaska.pipeline),
                           lab = "lab", depth = alaska.pipeline$lab.defect),
                data.frame(id = 1:nrow(alaska.pipeline),
                           lab="field", depth = alaska.pipeline$field.defect))
alaska$lab <- as.factor(alaska$lab)
lm(depth~lab, data = alaska)
```

```
##
## Call:
## lm(formula = depth ~ lab, data = alaska)
##
## Coefficients:
## (Intercept)       lablab
##       33.58         5.52
```

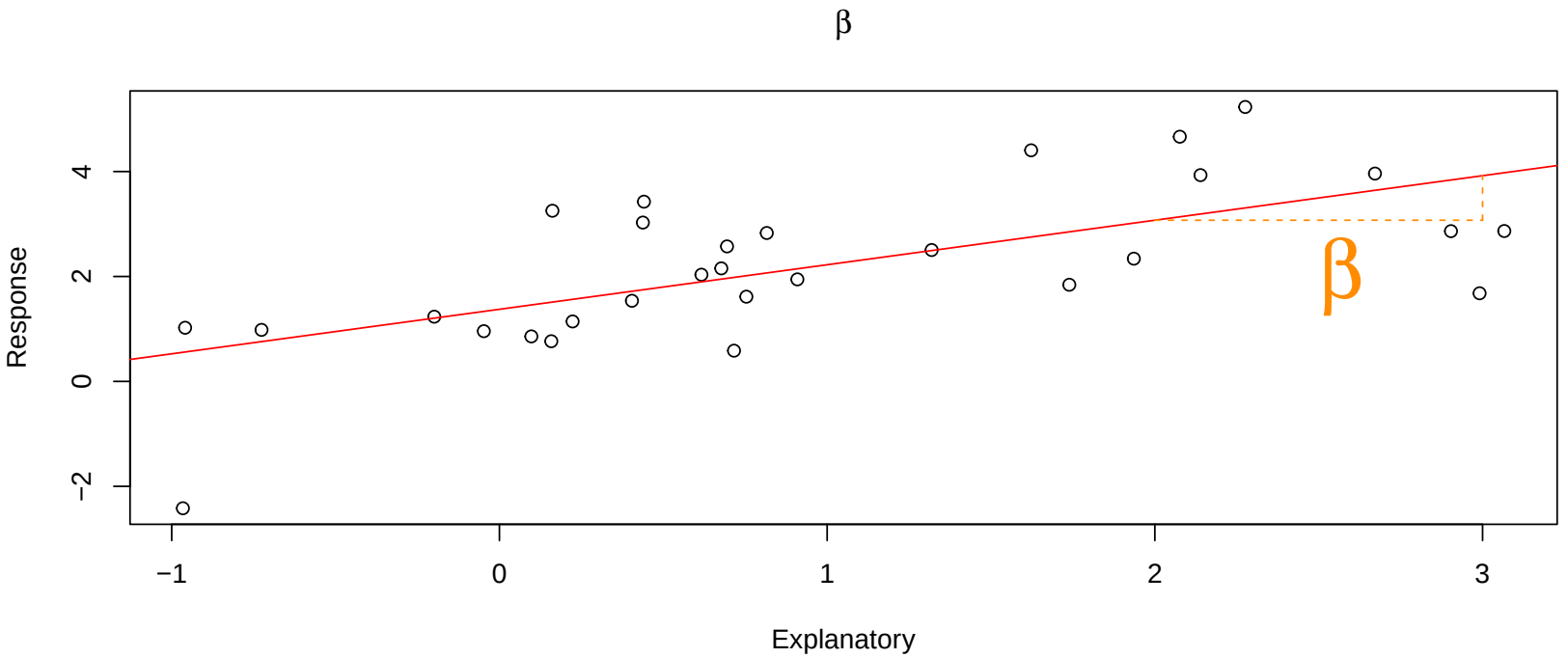
Examples of linear models: continuous x_i



$$y_i = \alpha + x_i\beta + \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- α: the intercept is the value of y_i where $x_i = 0$

Examples of linear models: continuous x_i



$$y_i = \alpha + x_i\beta + \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- α: the intercept is the value of y_i where $x_i = 0$
- β: the slope is the change in y_i for a unit increase in x_i

Alaska pipeline: lab and field defects

$$
\begin{aligned}
y_i &= \alpha + x_i\beta + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \\
lab.defect_i &= \alpha + field.defect_i * \beta + left.over.stuff
\end{aligned}
\tag{9}
$$

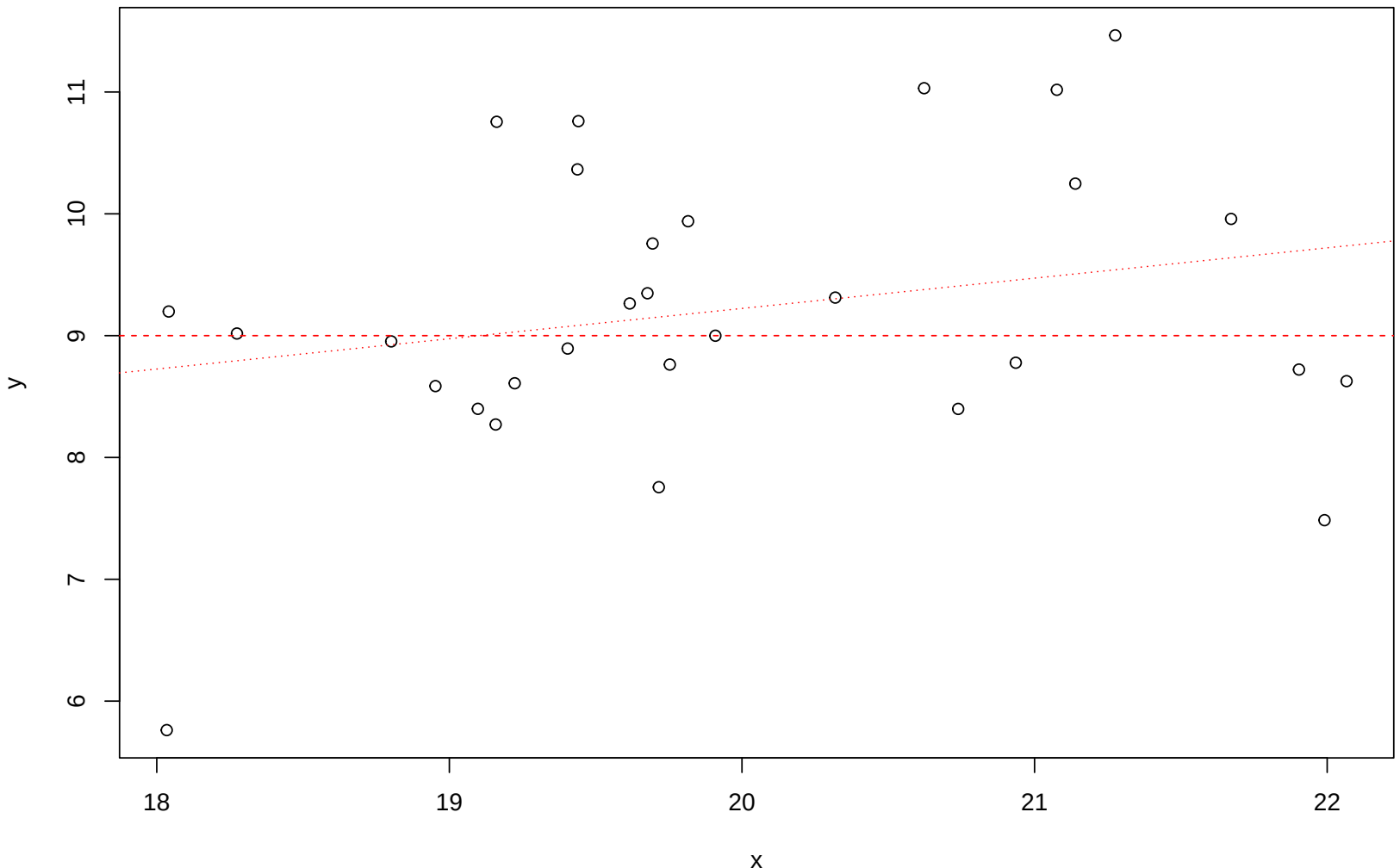
Alaska pipeline: lab and field defects

$$
\begin{aligned}
y_i &= \alpha + x_i\beta + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \\
lab.defect_i &= \alpha + field.defect_i * \beta + left.over.stuff
\end{aligned}
\tag{9}
$$

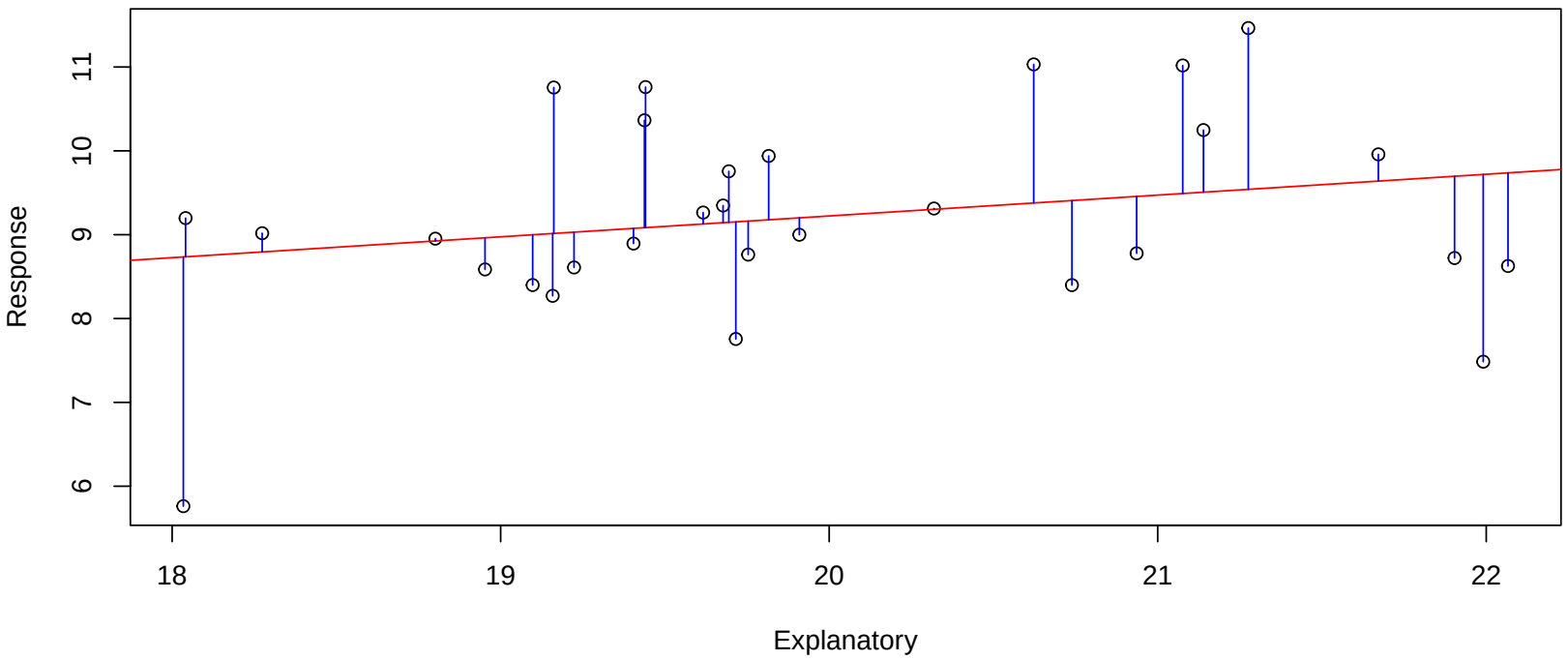
```
lm(lab.defect~field.defect, data = alaska.pipeline)
```

```
##
## Call:
## lm(formula = lab.defect ~ field.defect, data = alaska.pipeline)
##
## Coefficients:
##  (Intercept)  field.defect
##       -1.967         1.223
```

What is the best line?



How good is the line?



Distance from model (line) to data: "error" ϵ_i

Least squares estimation

Minimize the sum of squared residuals:

$$RSS = \sum_{i=1}^{n} \epsilon_i^2 \tag{10}$$

which is the same to maximizing the normal likelihood!

- $\hat{\alpha} = \frac{1}{N}(\sum y_i - \hat{\beta} \sum x_i)$
- $\hat{\beta} = \frac{\sum(x_i - \sum \frac{x_i}{N})(y_i - \sum \frac{y_i}{N})}{\sum(x_i - \sum \frac{x_i}{N})^2}$
- $\hat{\sigma^2} = \frac{1}{N-1} \sum (y_i - (\alpha + x_i \beta))^2$
- $\hat{\mu}_i = \hat{\alpha} + x_i \hat{\beta}$

Our model

$$y_i = \alpha + \beta x_i + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{11}$$

- y_i: our data
- α, β describe our line
- ϵ_i quantifies distance to the model

Assumptions

We make some critical assumptions here

1) the relationship between y_i and x_i
2) distribution of the errors

Other assumptions for the Errors

$$\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \tag{12}$$

- are normally distributed
- have constant variance ("Homoscedastcity")
- are independent
- no outliers

Assumptions for the errors

all the errors together tell us how good the line is

- this is the same as finding the line by maximum likelihood estimation 😄

$$y_i \sim \mathcal{N}(\alpha + x_i\beta, \sigma^2) \tag{13}$$

Summary simple linear models

- includes t-test, anova (analysis of variance) and regression
- all use the same mathy bits (and model)
- **interpretation** depends on the type of **variable**
- GLMs take the same form

Summary

- Normal-distribution and t-distribution
- Simple linear regression: one covariate
- Least squares estimation
- Difference in LMs with categorical or continuous explanatory variables
  - Categorical: intercept/mean parameter
  - Continuous: slope parameter
- Fortunately we have the `lm()` function in R!
- More on assumptions checking tomorrow

Summary (2)

t-test: linear model with categorical covariate of two groups
ANOVA: linear model with categorical covariate of multiple groups
ANCOVA: linear model with categorical and continuous covariate (interaction)