DIVING INTO THE LINEAR MODEL

DEEPER INTO THE WORLD OF LINEAR MODELS

1. Are groups different: T, F, and Multiple Comparisons

2. Linear Regression

3. Nonlinear Regression

THE KNEE'S THE THING!

- Test if shining light behind knees or on eyes helped jet lag
- Also a control group with no light
- n=7

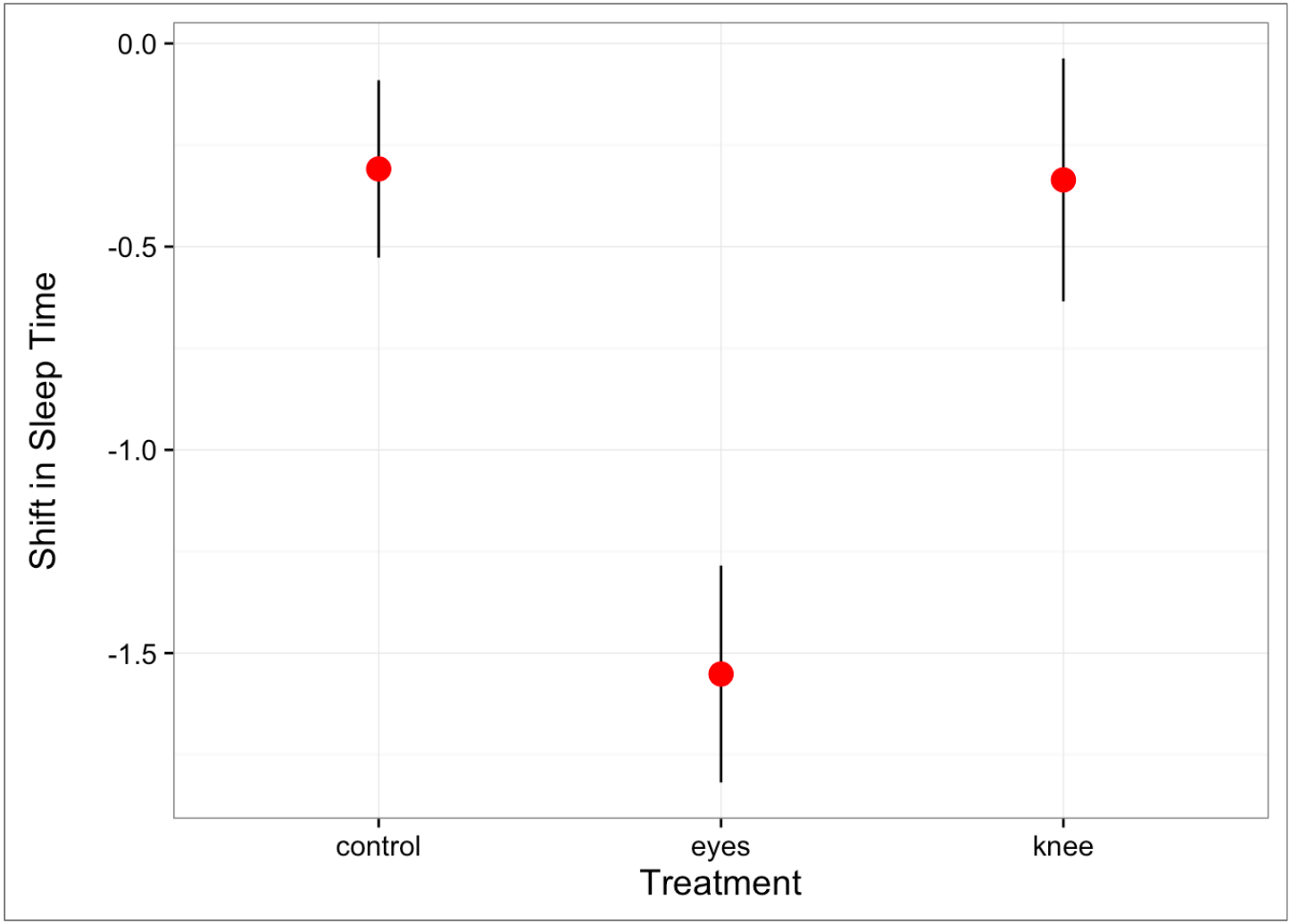
MANY QUESTIONS

1. Are groups different from 0?

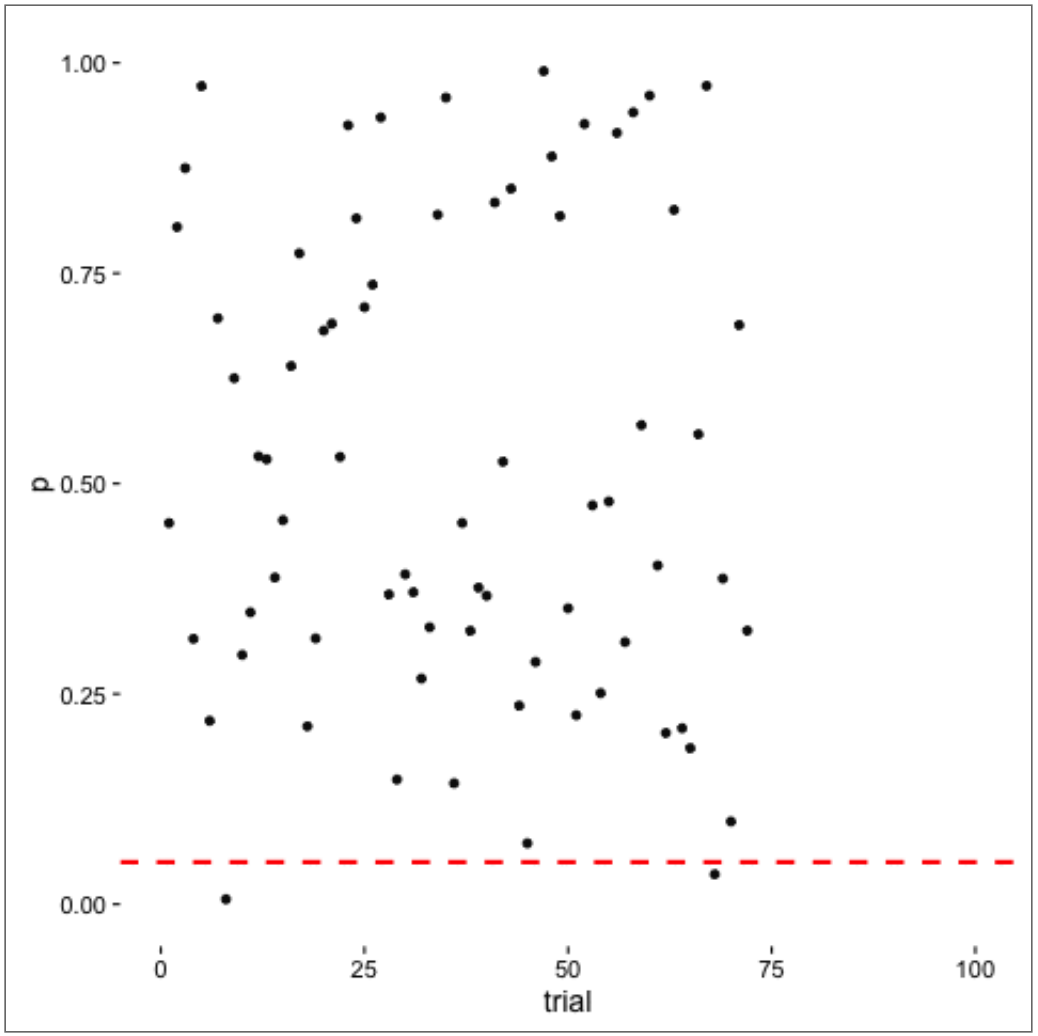
2. Are groups different from each other?

3. Are groups a meaningful explanatory variable?

WHICH OF THESE GROUPS DIFFER: A JOB FOR T



THE PROBLEM OF MULTIPLE COMPARISONS



SOLUTIONS TO MULTIPLE COMPARISONS?

1. Ignore it - a test is a test

2. Lower your α given m = # of comparisons
    - Bonferroni α/m
    - False Discovery Rate $k\alpha/m$ where k is rank of test
3. Other multiple comparinson correction
    - Tukey's Honestly Significant Difference

NO CORRECTION

```
pairwise.t.test(jetlag$shift, jetlag$treatment,
                p.adjust.method="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  jetlag$shift and jetlag$treatment
##
##      control eyes
## eyes 0.0029  -
## knee 0.9418  0.0044
##
## P value adjustment method: none
```

BONFERRONI CORRECTIONS

```
pairwise.t.test(jetlag$shift, jetlag$treatment,
                p.adjust.method="bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  jetlag$shift and jetlag$treatment
##
##      control eyes
## eyes 0.0088  -
## knee 1.0000  0.0132
##
## P value adjustment method: bonferroni
```

FDR

```
pairwise.t.test(jetlag$shift, jetlag$treatment,
                p.adjust.method="fdr")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  jetlag$shift and jetlag$treatment
##
##      control eyes
## eyes 0.0066  -
## knee 0.9418  0.0066
##
## P value adjustment method: fdr
```

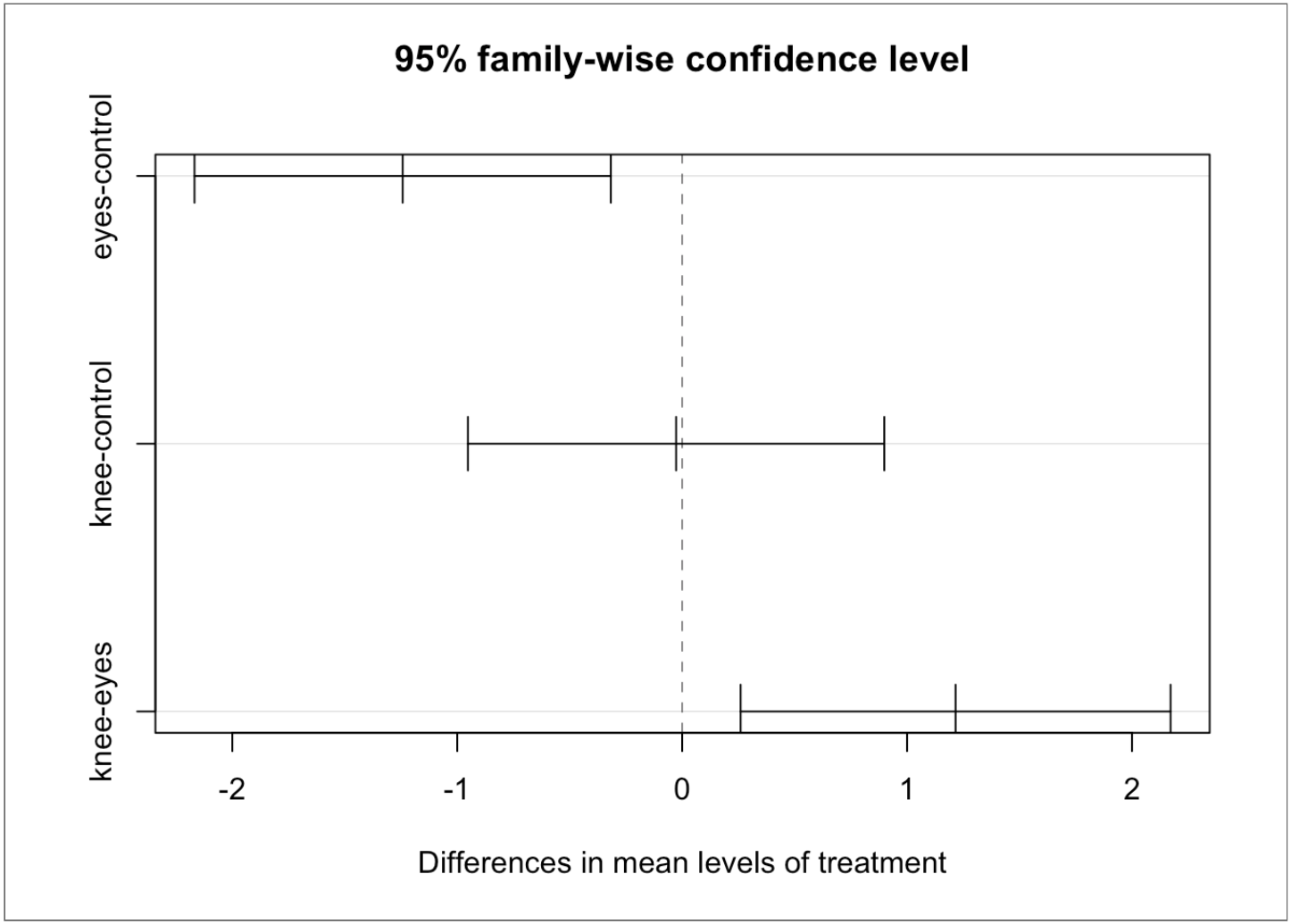
TUKEY TEST

```
jet_aov <- aov(shift ~ treatment, data=jetlag)
```

```
TukeyHSD(jet_aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = shift ~ treatment, data = jetlag)
##
## $treatment
##                     diff        lwr        upr     p adj
## eyes-control -1.24267857 -2.1682364 -0.3171207 0.0078656
## knee-control -0.02696429 -0.9525222  0.8985936 0.9969851
## knee-eyes     1.21571429  0.2598022  2.1716263 0.0116776
```

VISUALIZING COMPARISONS



DEEPER INTO THE WORLD OF LINEAR MODELS

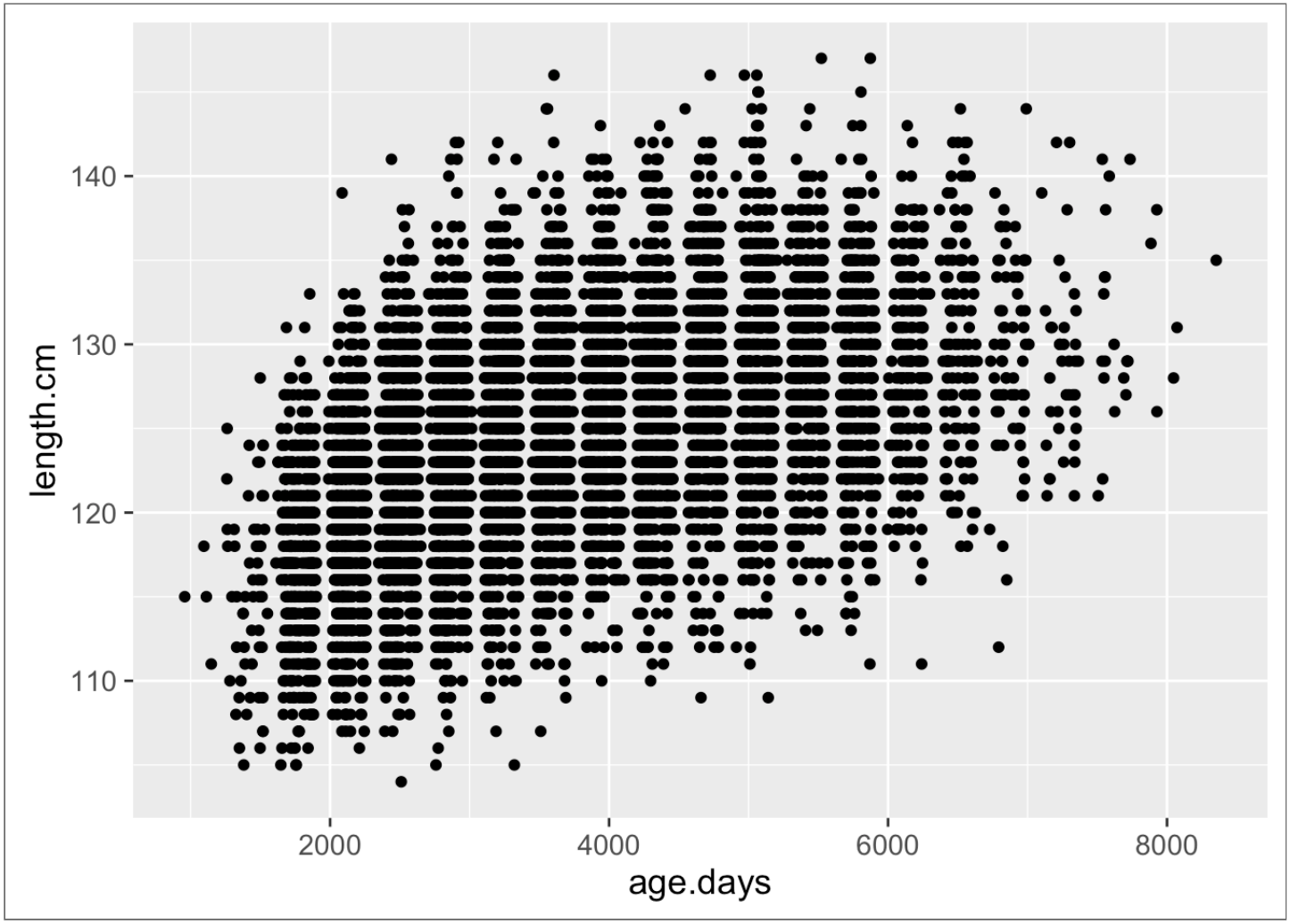
1. Are groups different: T, F, and Multiple Comparisons

2. Linear Regression

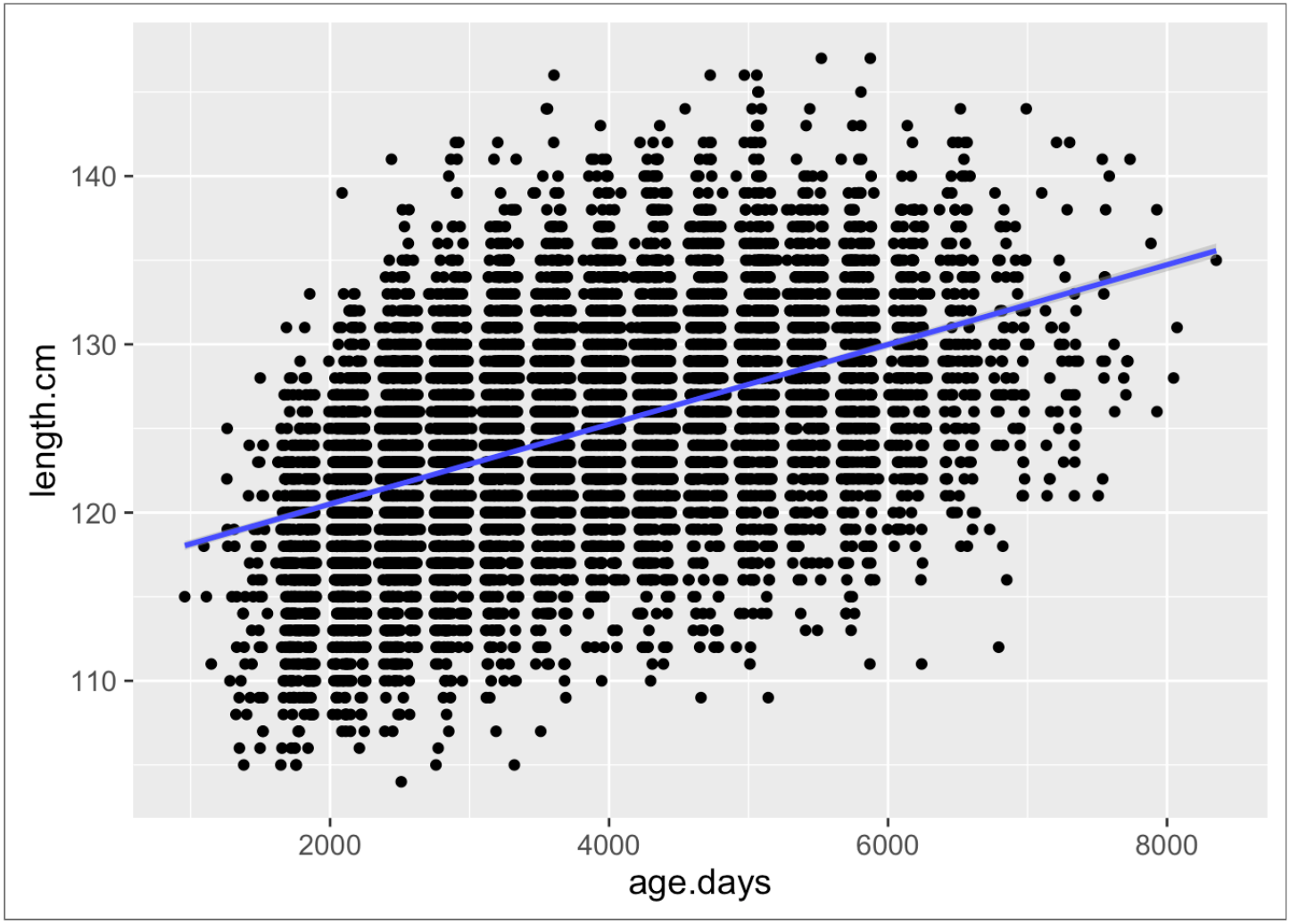
3. Nonlinear Regression

AGING SEALS

ARE OLDER SEALS BIGGER?



ARE OLDER SEALS BIGGER?



THE STEPS OF STATISTICAL MODELING

1. What is your question?
2. What model of the world matches your question?
3. Build a test
4. Evaluate test assumptions
5. Evaluate test results
6. Visualize

OUR QUESTION AND MODEL

Question 1: Are older seals bigger?

Question 2: How much do seals grow per day?

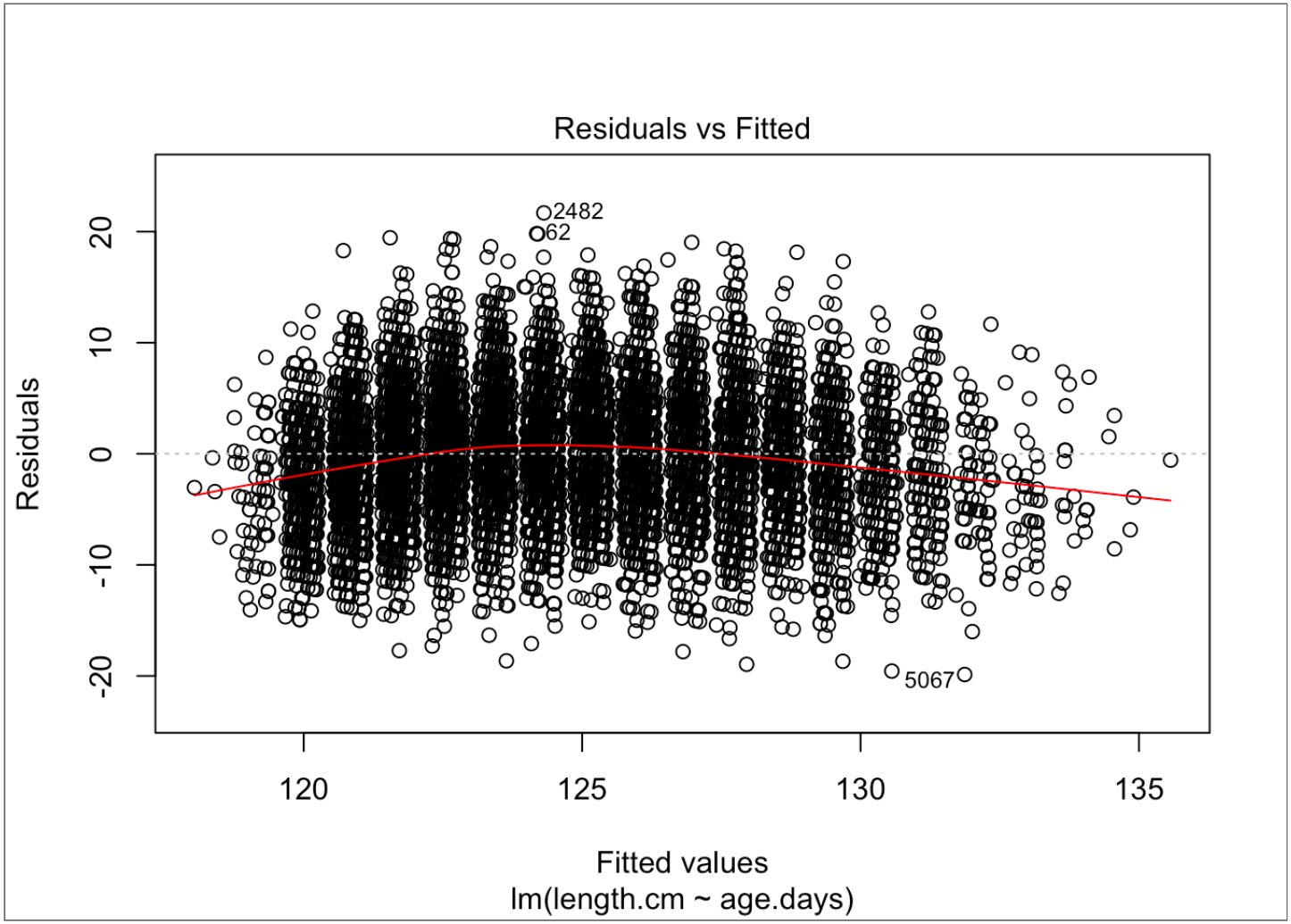
$$Length_i = \beta_0 + \beta_1 Age_i + \epsilon_i$$

MORE LINEAR MODELS

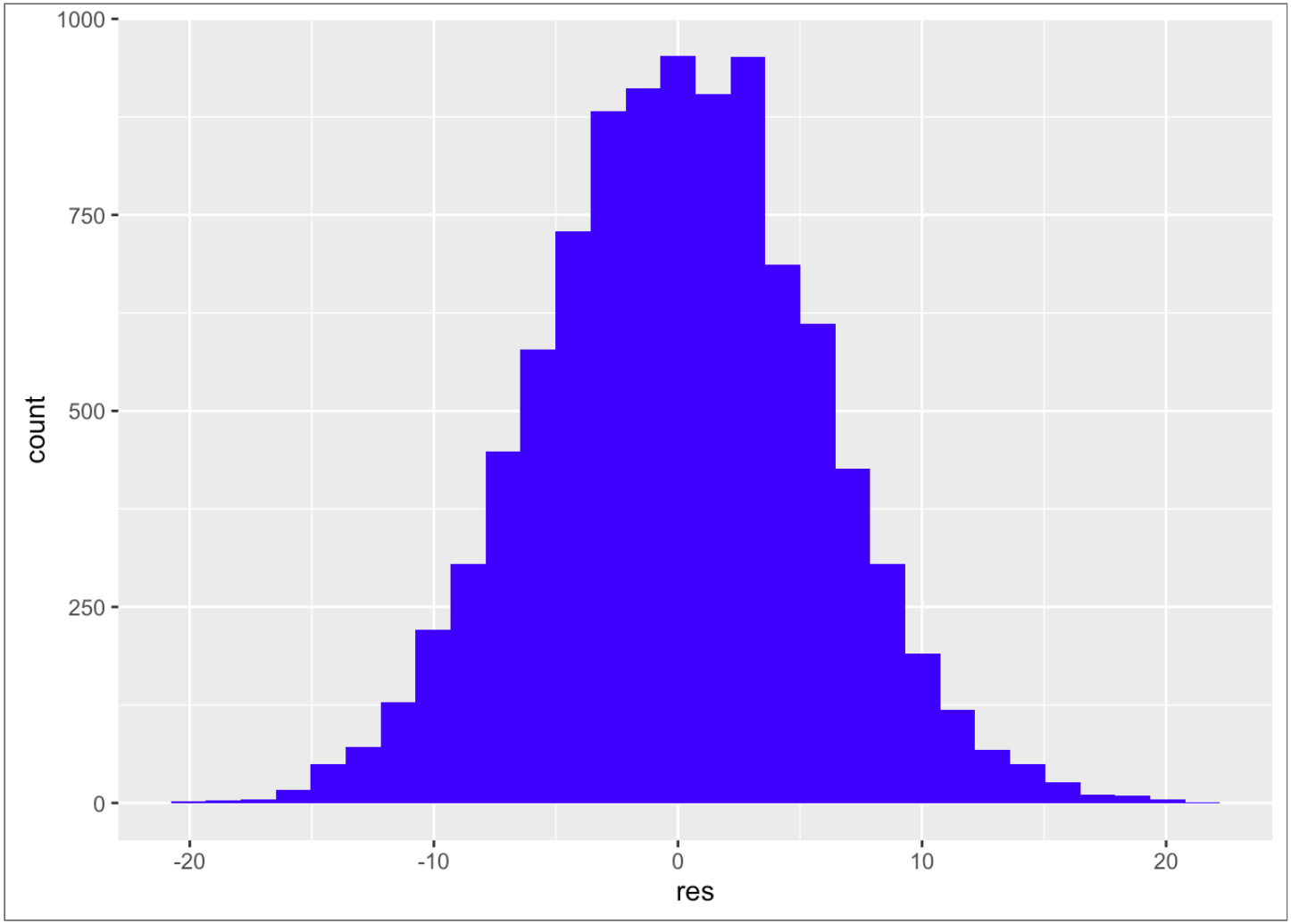
$$Length_i = \beta_0 + \beta_1 Age_i$$

```
seal_mod_linear <- lm(length.cm ~ age.days, data=seals)
```

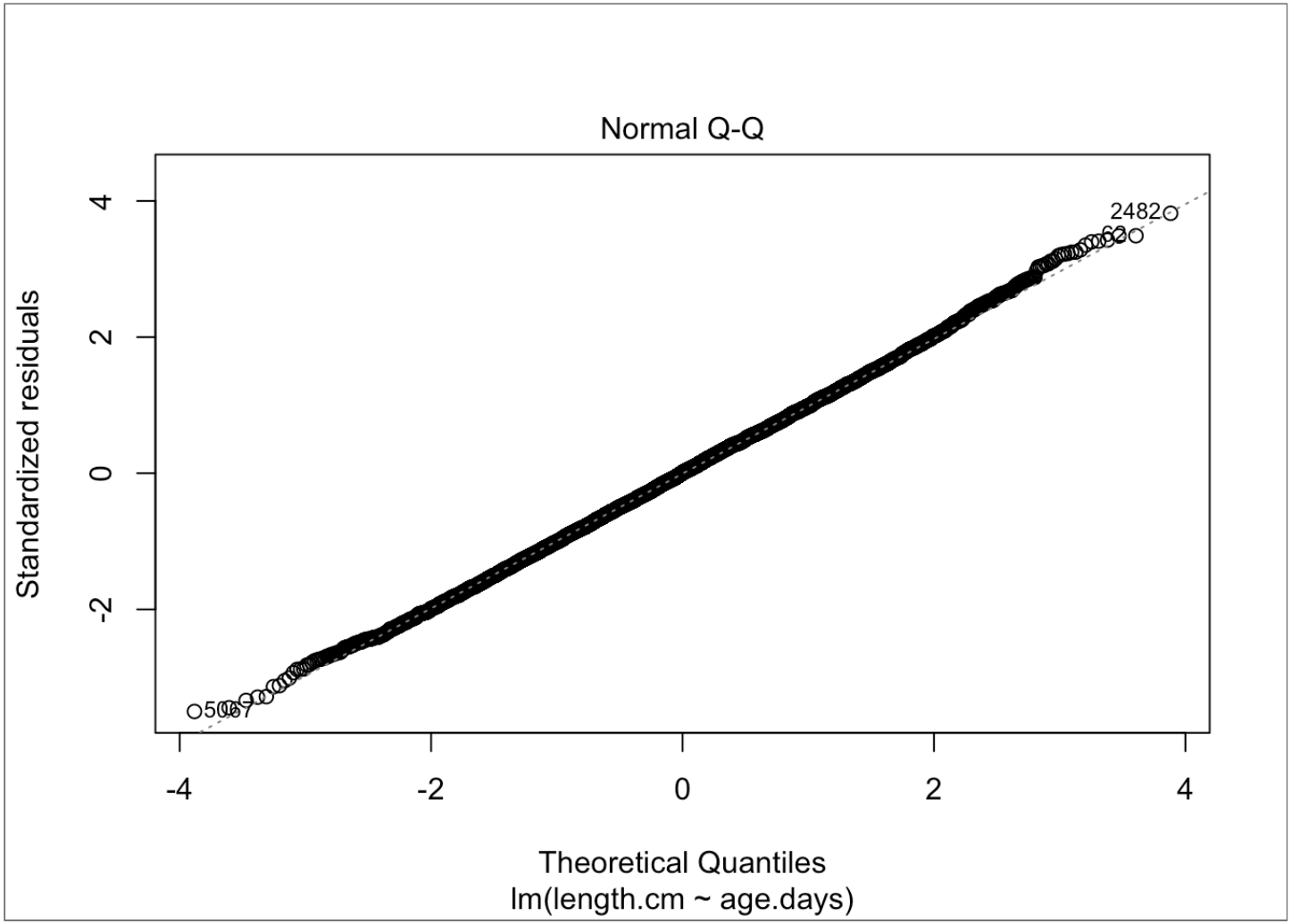
WAS THIS LINEAR? FITTED VERSUS RESIDUALS



TESTING OUR ERROR GENERATING PROCESS



QQS EVERYWHERE



IS OUR MODEL ANY GOOD? F-THAT!

```
anova(seal_mod_linear)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age.days	1	90861.81	90861.80648	2815.841	0
Residuals	9663	311806.52	32.26809	NA	NA

T FOR COEFFICIENTS

```
summary(seal_mod_linear)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	115.7667505	0.1763751	656.3668	0
age.days	0.0023706	0.0000447	53.0645	0

HOW MUCH VARIATION DOES OUR MODEL EXPLAIN?

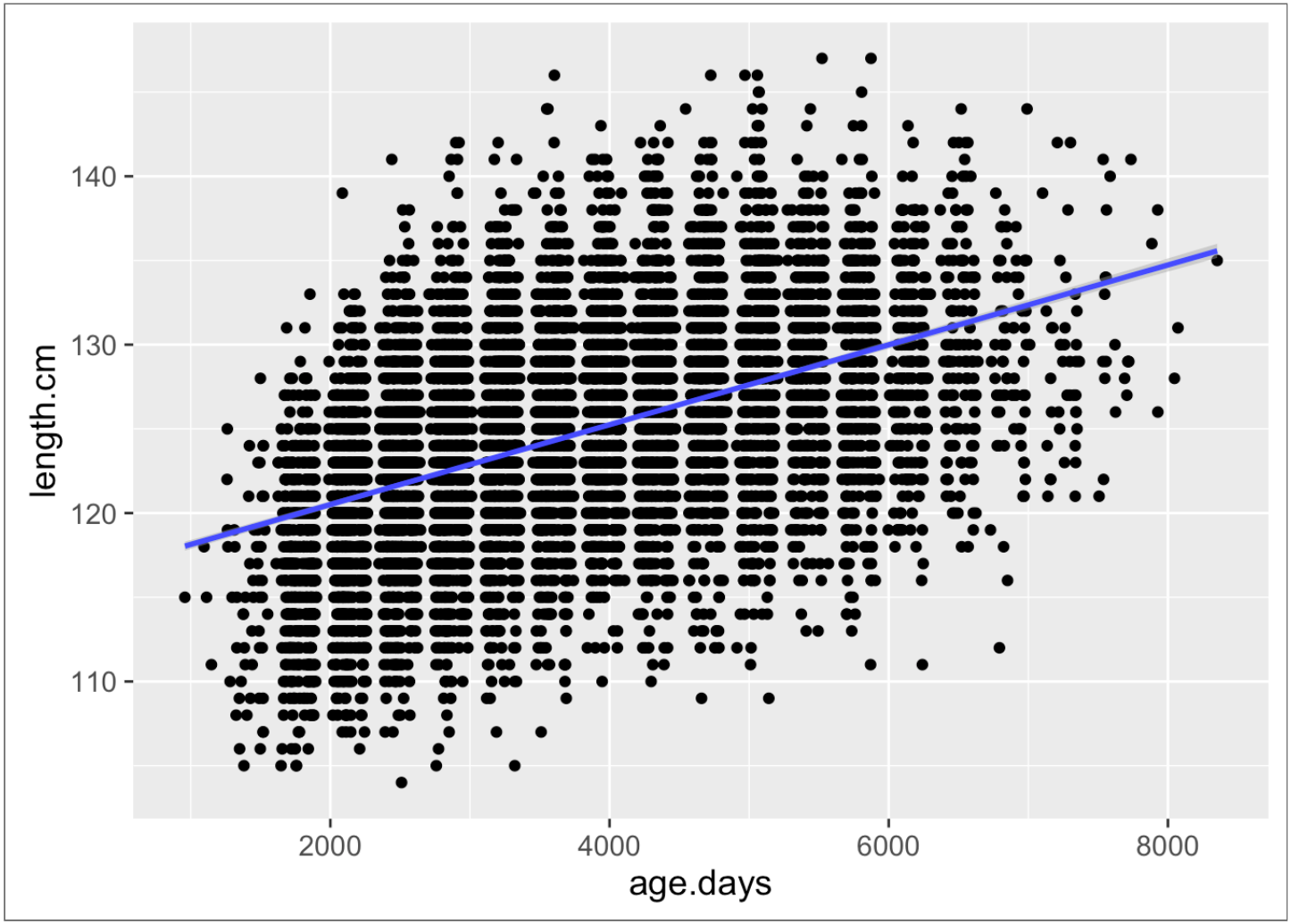
$1 - Var_{residual}/Var_{total}$

$R^2 = 0.2256493$

VISUALIZATION

```
seal_base <- ggplot(seals, aes(x=age.days, y=length.cm)) +
  geom_point() +
  theme_grey(base_size=14) +
  stat_smooth(method="lm")
```

VISUALIZATION



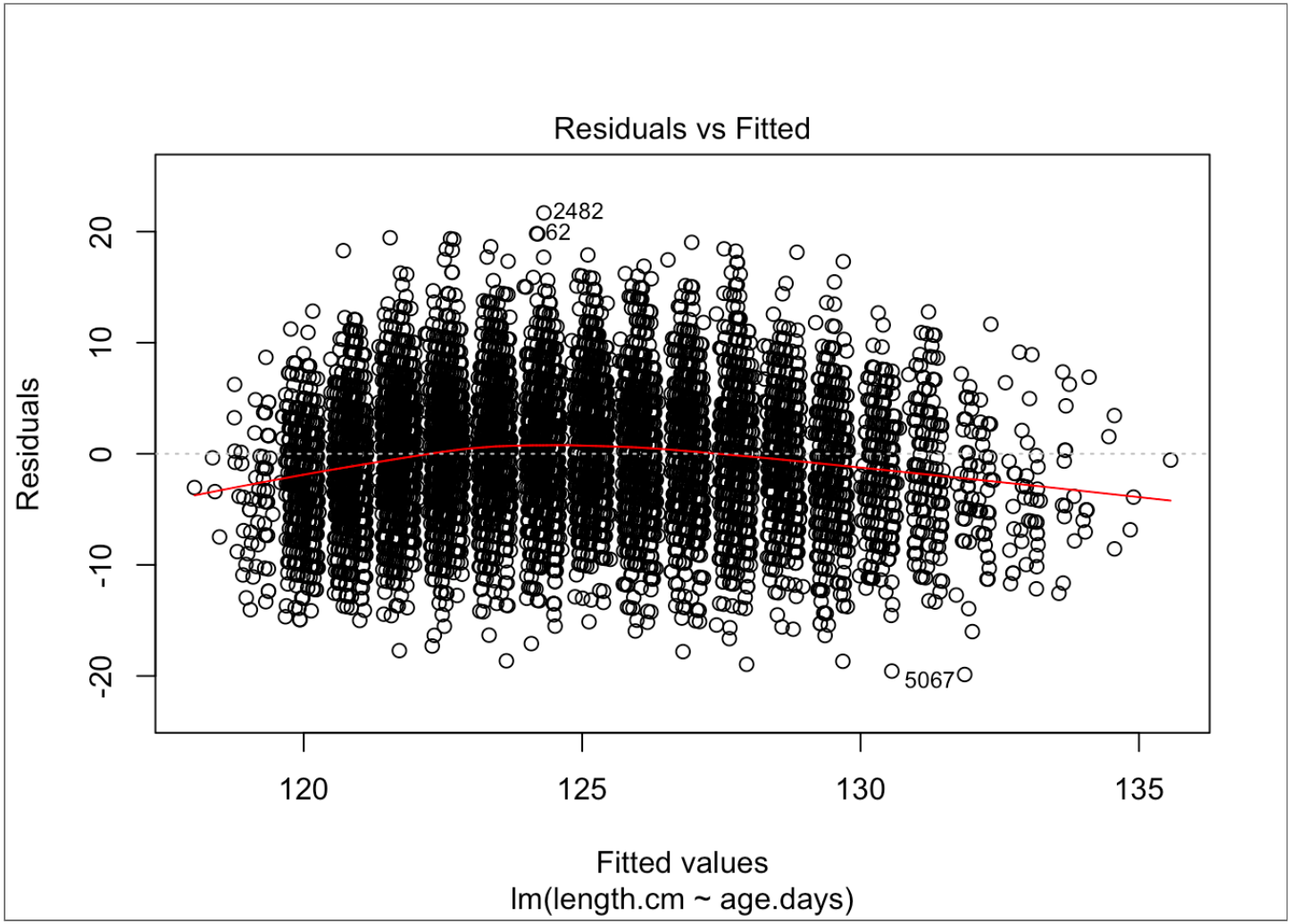
DEEPER INTO THE WORLD OF LINEAR MODELS

1. Are groups different: T, F, and Multiple Comparisons

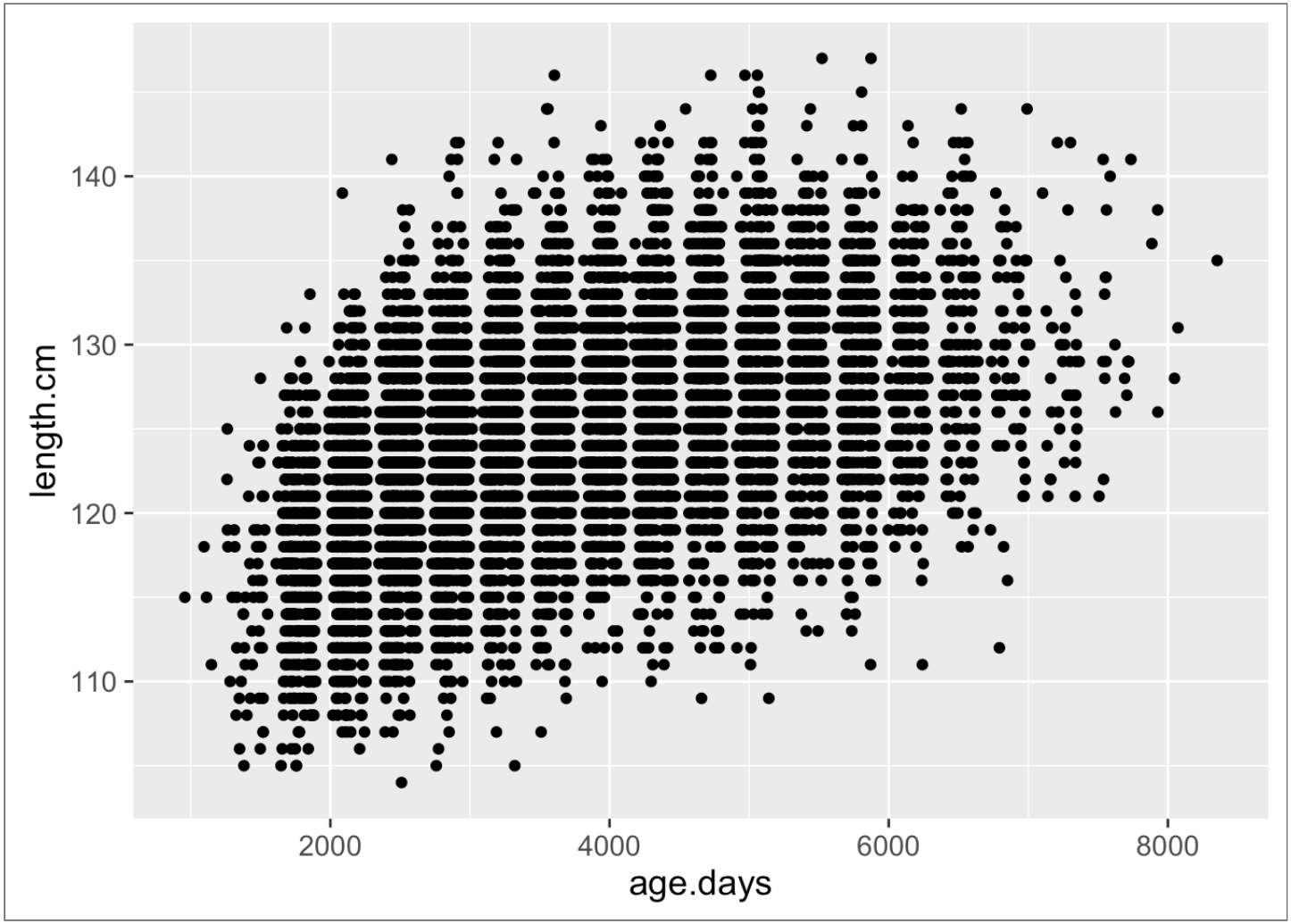
2. Linear Regression

3. Nonlinear Regression

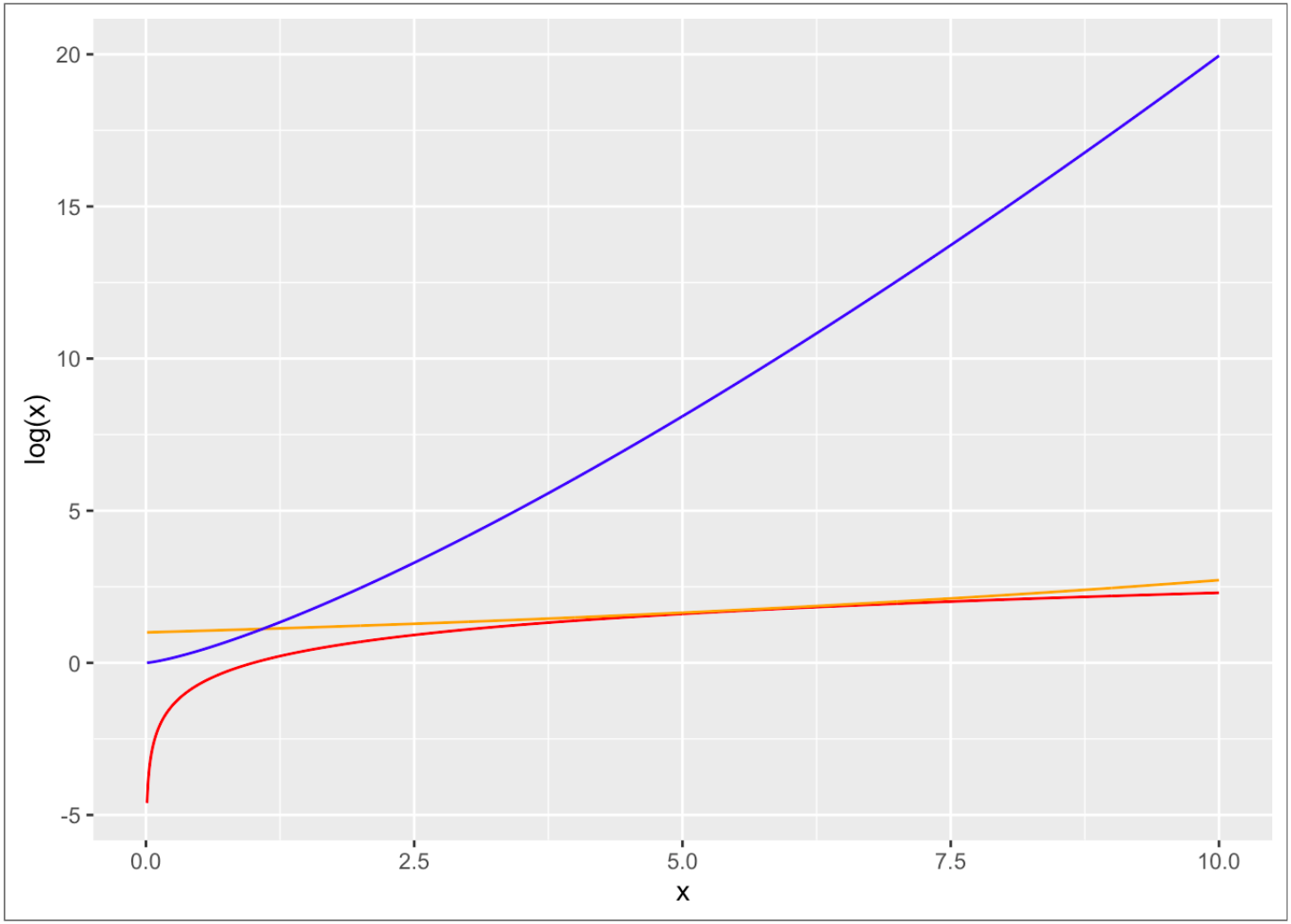
RESIDUALS V. FITTED FISHY…



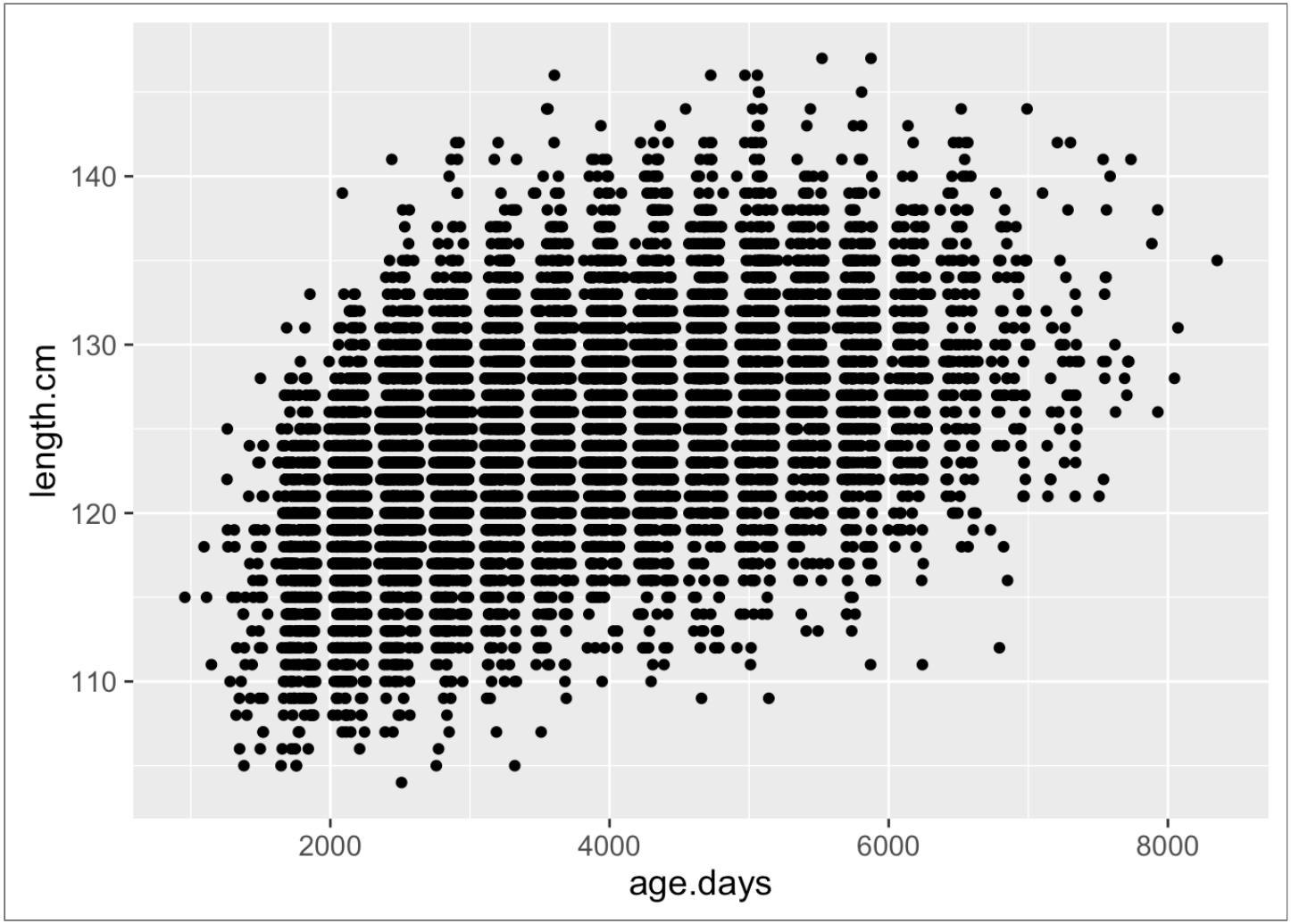
A DIFFERENT DATA GENERATING PROCESS?



POSSIBLE NONLINEAR DATA GENERATING PROCESSES



WHAT SHAPE IS THE DATA GENERATING PROCESS?

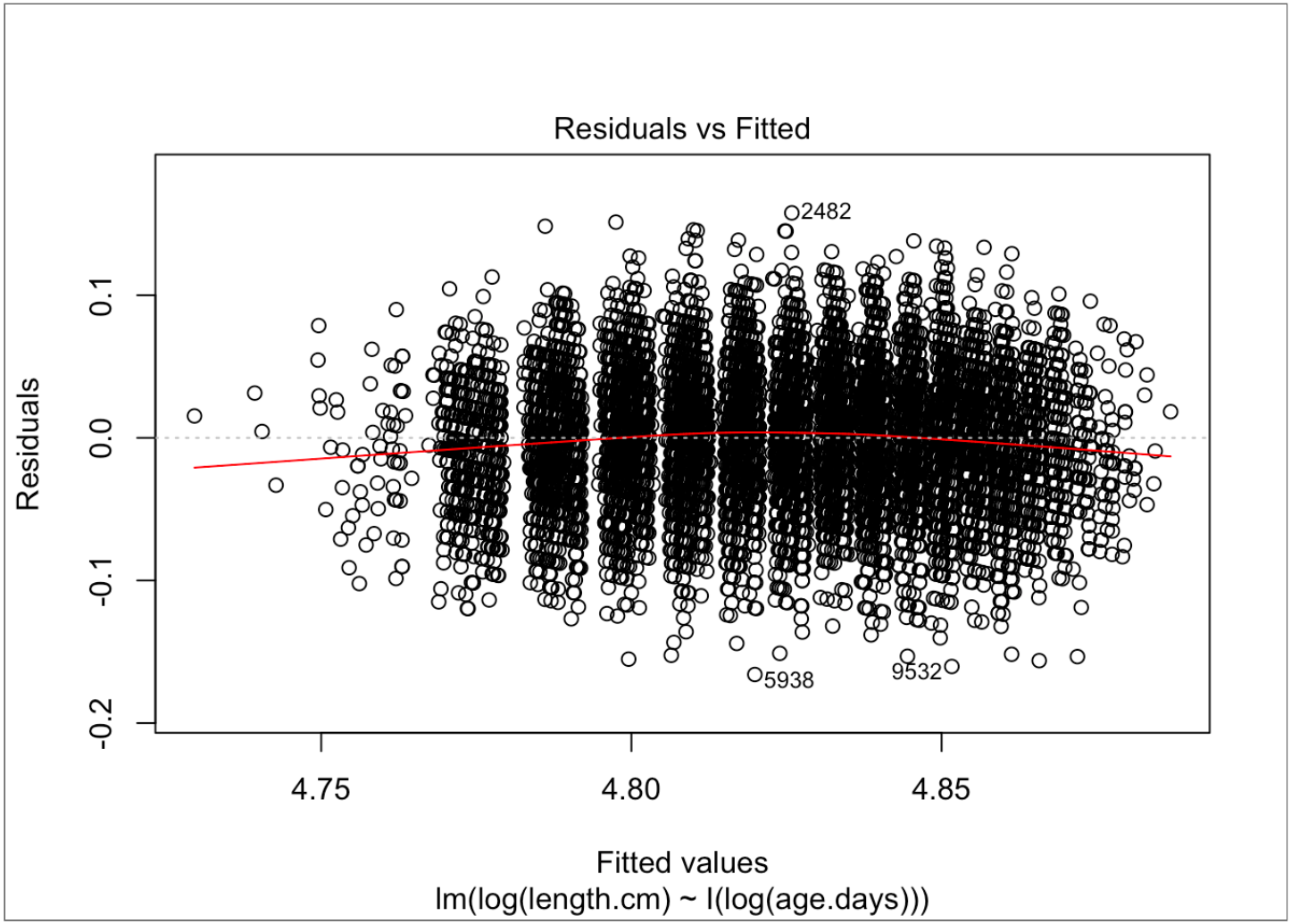


A NEW MODEL

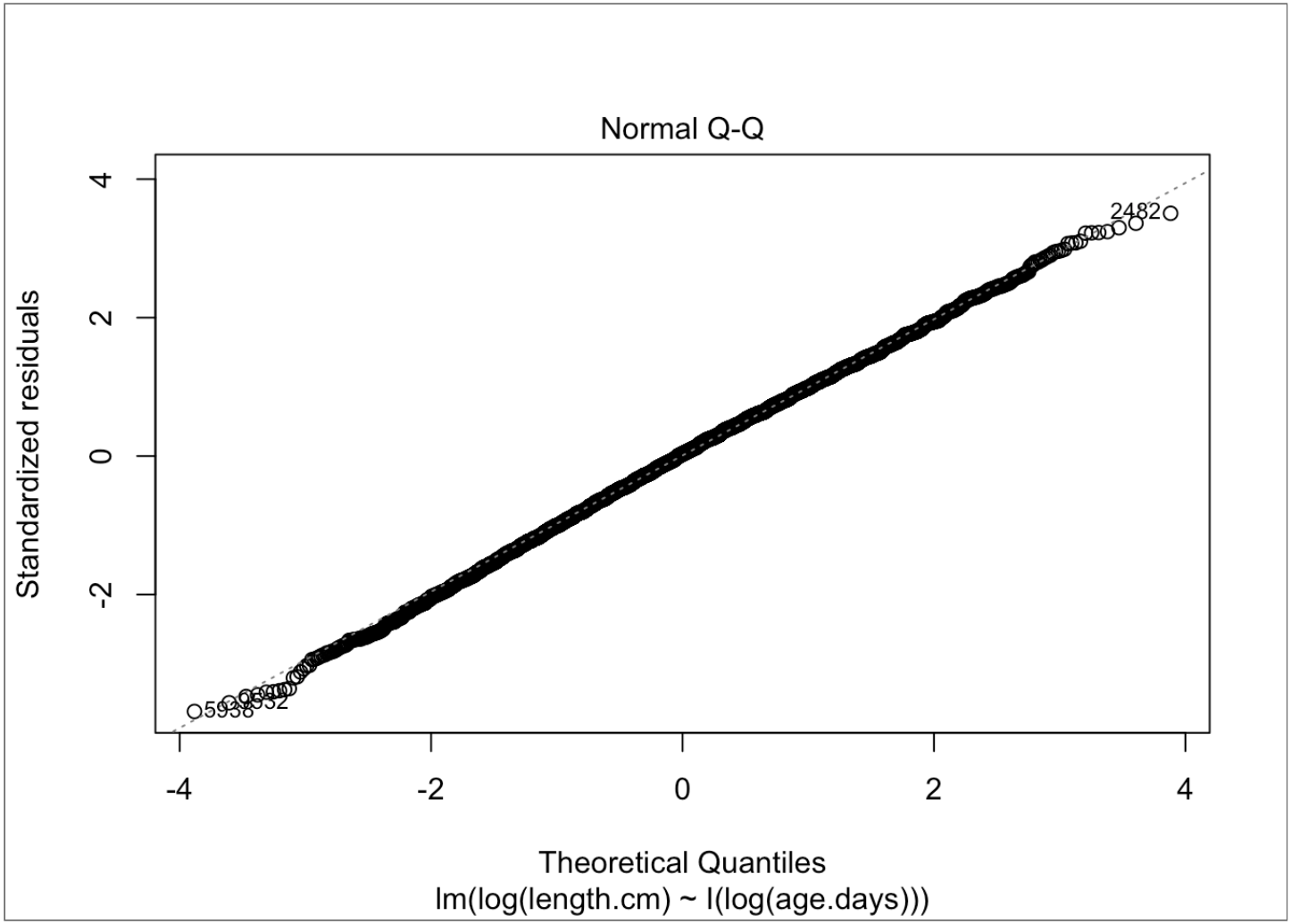
$$y = \beta_0 + \beta_1 log(x) + \epsilon$$

```
seal_mod_log <- lm(log(length.cm) ~ I(log(age.days)), data=seals)
```

DID IT FIX THE RESIDUALS PROBLEM?



STILL MUST EVALUATE ERROR GENERATING PROCESS



HOW MUCH VARIATION DOES OUR MODEL EXPLAIN?

$1 - Var_{residual}/Var_{total}$

$R^2 = 0.2500709$

VISUALIZATION

```
seal_base +
  stat_smooth(method="lm", formula=y ~ I(log(x)))
```