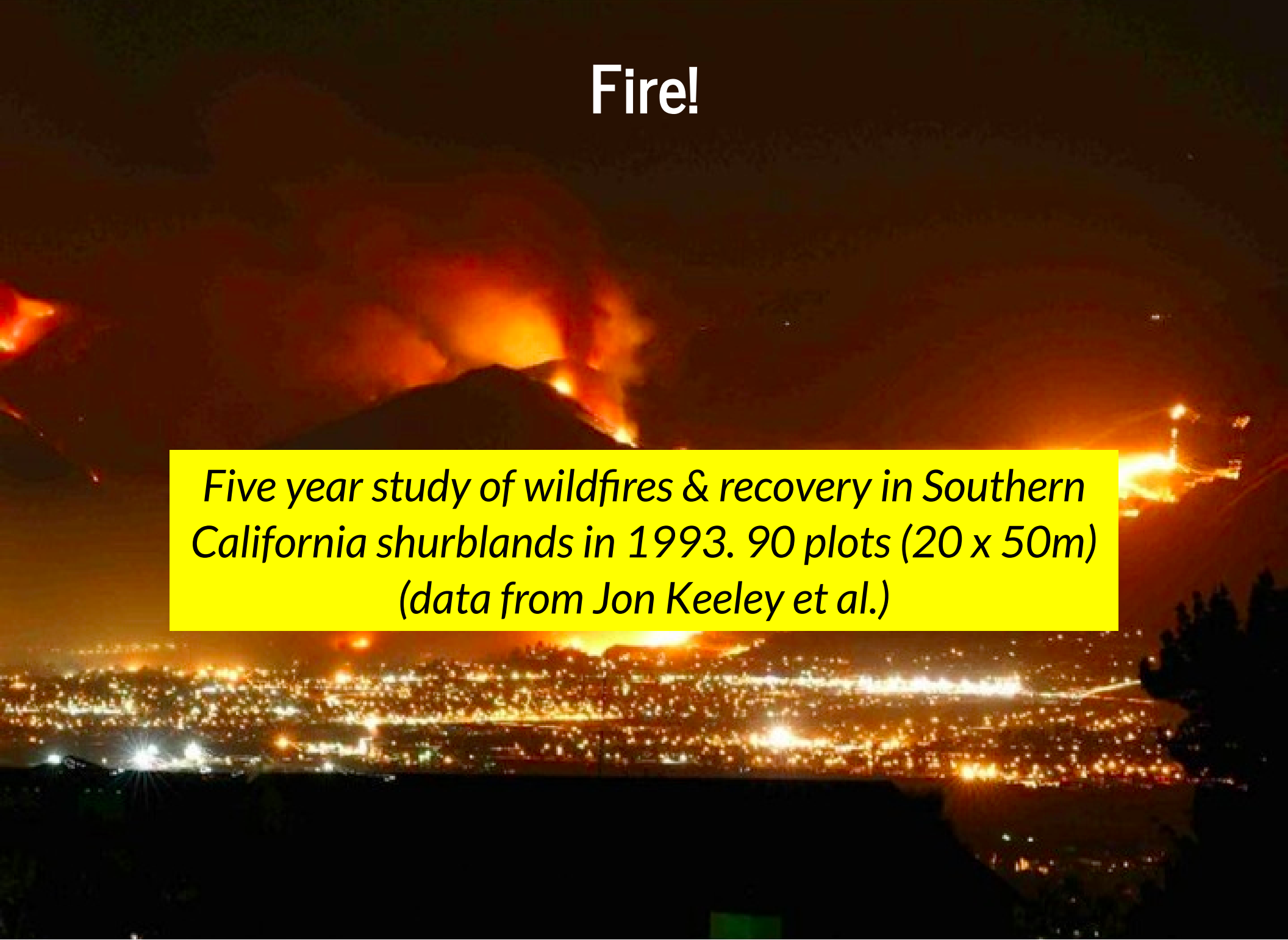
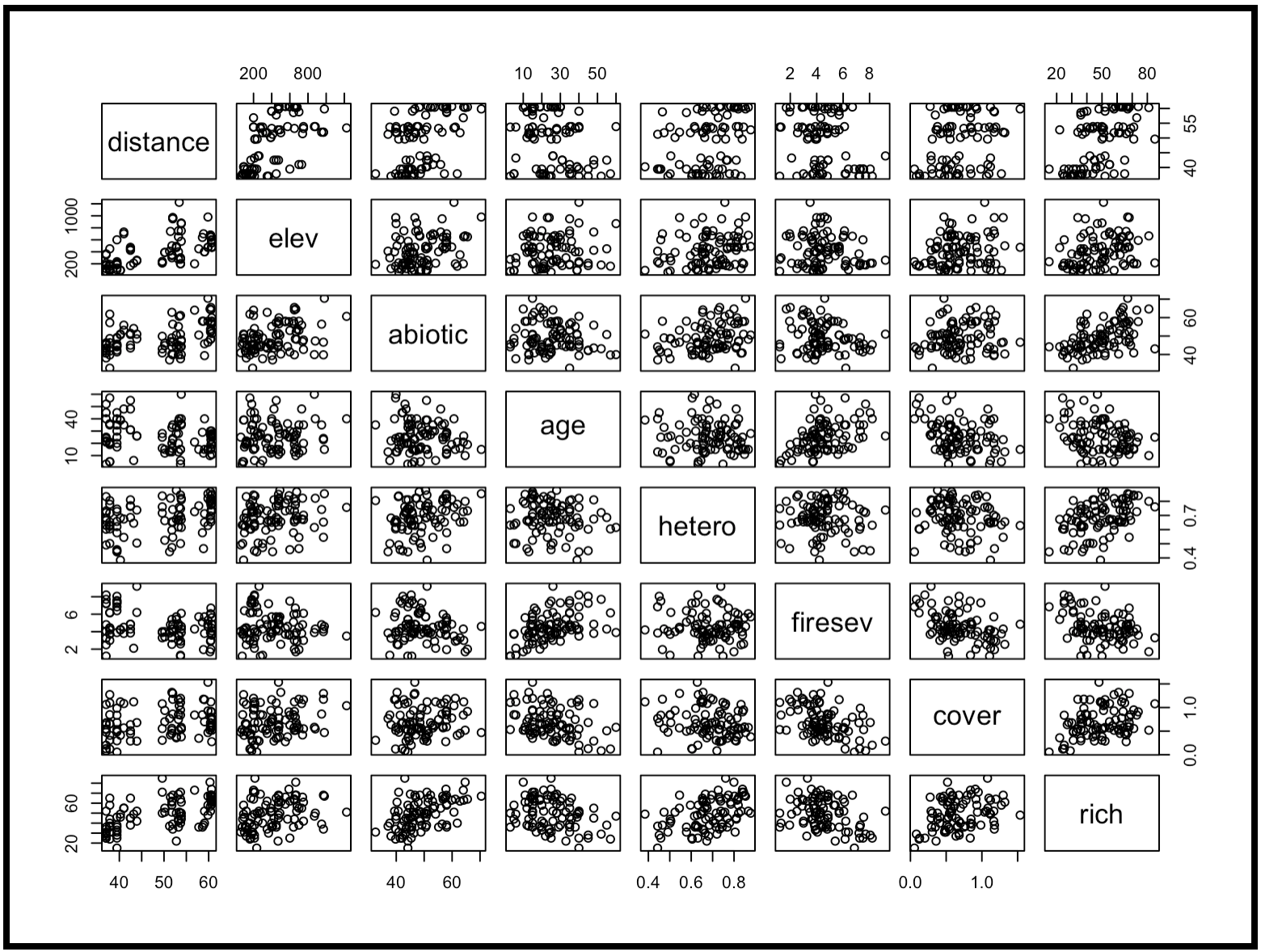
A World of Many Predictors

Fire!



Five year study of wildfires & recovery in Southern California shurblands in 1993. 90 plots (20 x 50m) (data from Jon Keeley et al.)

Many Things may Influence Species Richness



Q: How does Fire Influence Species Richness?

Possible causes:

- Distance from a burned patch
- Fire Severity
- Habitat heterogeneity

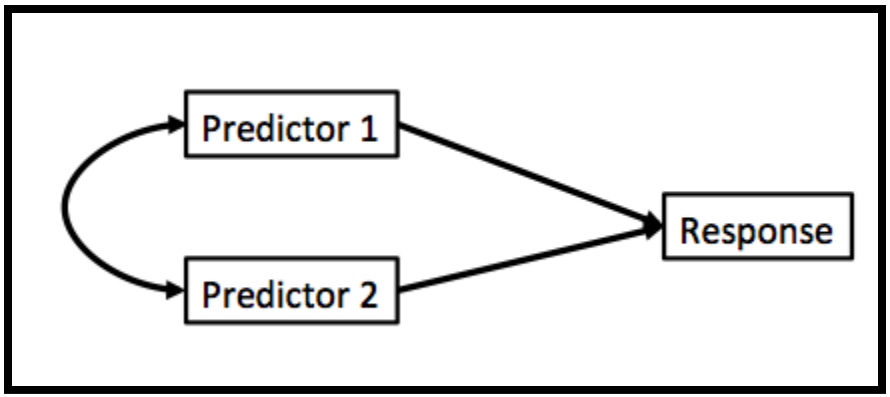
Our Model of How the World Works

$$Diversity_i = \beta_0 + \beta_1 Distance_i + \beta_2 Severity_i + \beta_3 Heteroge$$

$$\epsilon_i \sim N(0, \sigma)$$

- Linear additive data generating process
- Normal error generating process

Multiple Regression in 1 Slide



- We are looking at unique contribution of each predictor on a response

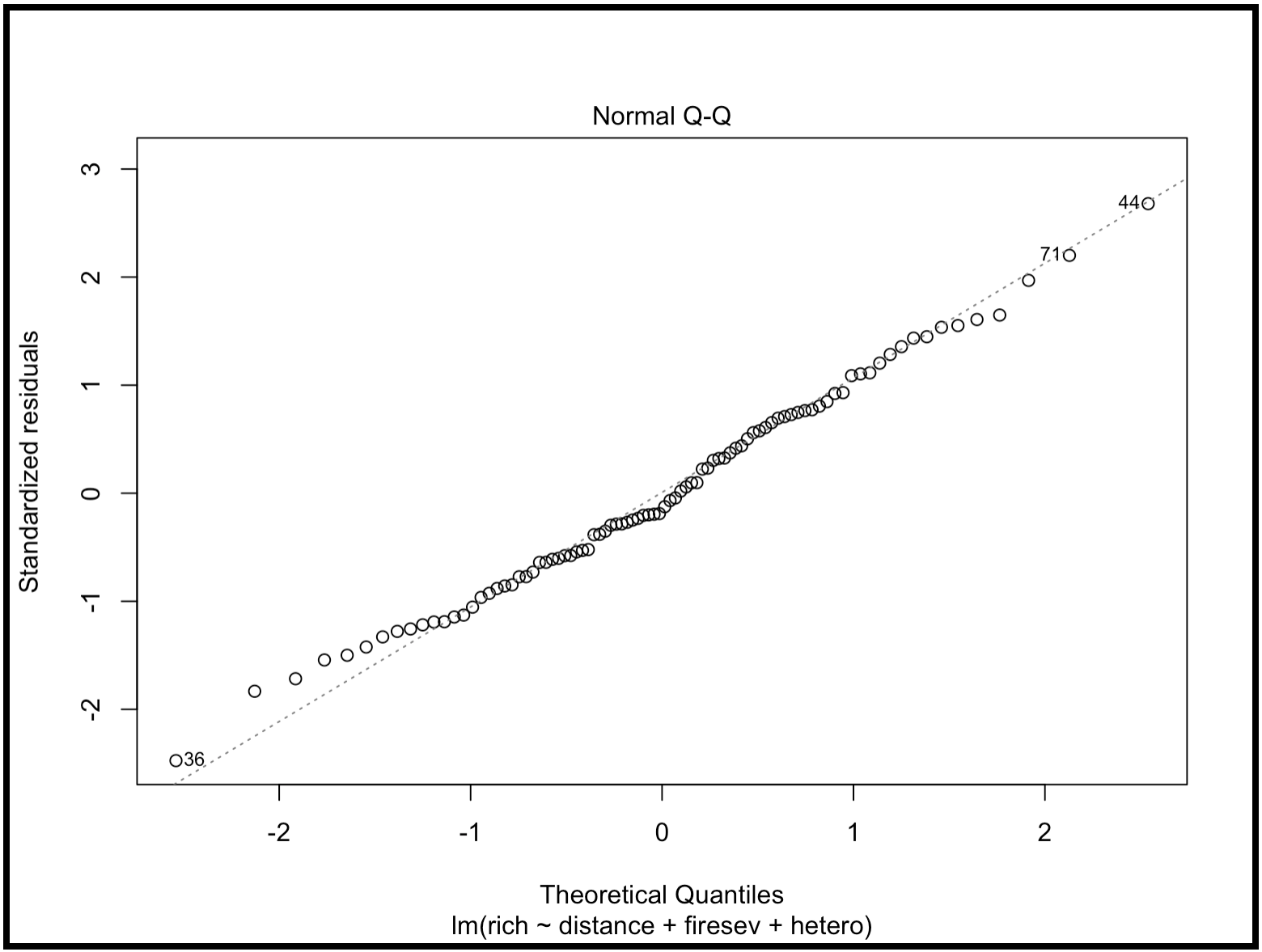
- We are **controlling** for covariation between predictors

Multiple Regression in R in 1 Slide

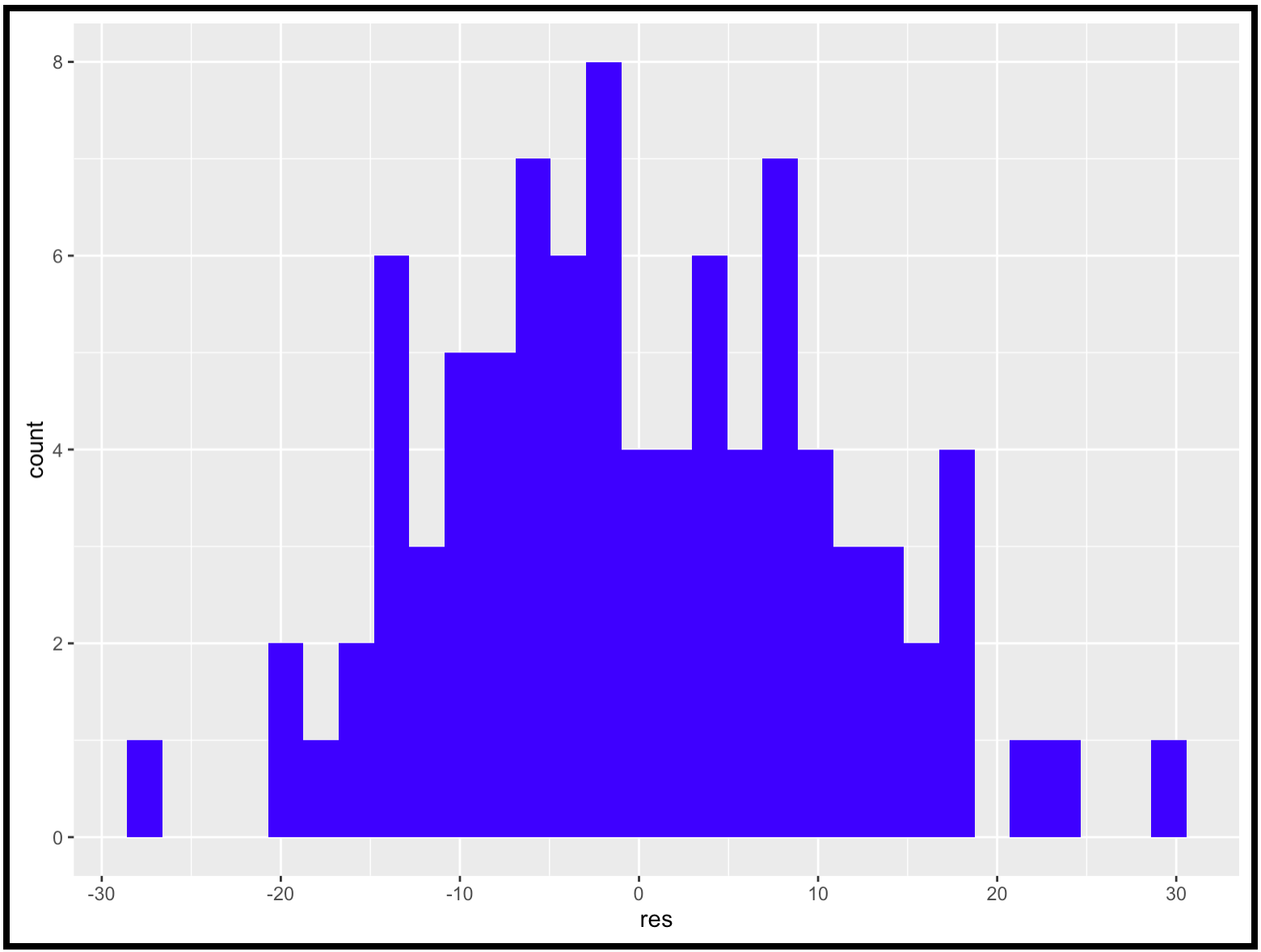
```
read.csv "./mlr/Keeley_rawdata_select4.csv"

lm                                    data=
```

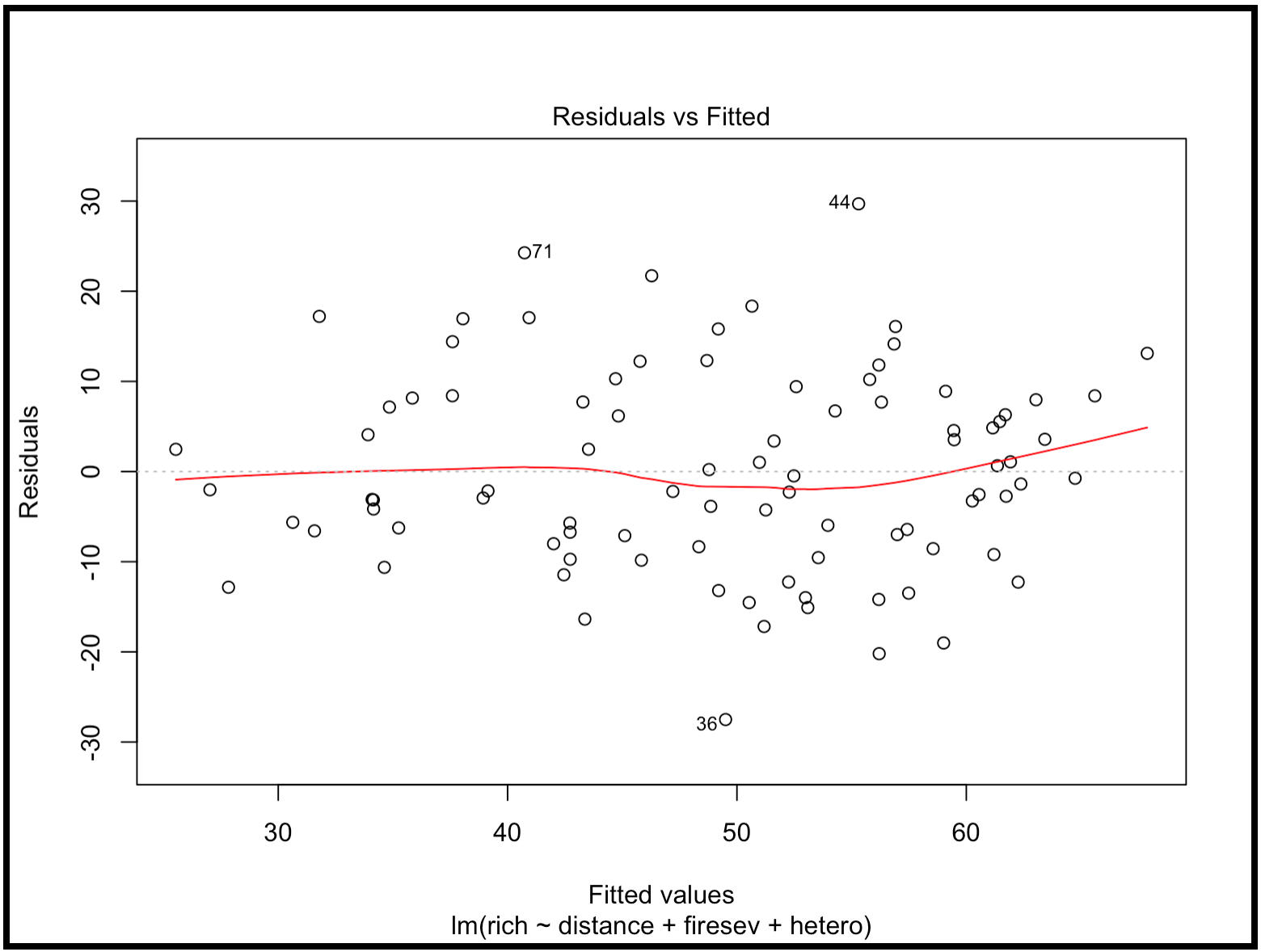
Still the Assumption of Normal Residuals



Still the Assumption of Normal Residuals

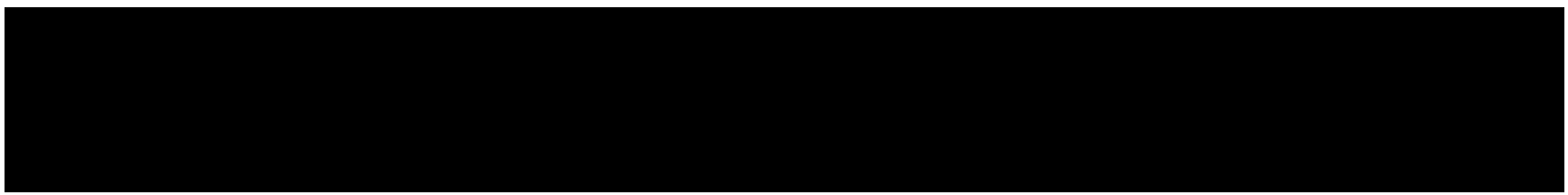


Still Must Inspect Fitted v. Residuals



New Assumption: Low Collinearity Between Predictors

- If predictors are highly correlated fit might be bad
- High correlation can lead to inflated error in parameters
- Above ~0.8 is bad, no unique information from variables



Second Test: Is Variance Inflated by Any Predictor?

Asks by what factor variance in parameter estimates is inflated by a predictor

```
library
vif
```

VIF > 5 or 10 can be problematic and indicate an unstable solution.

Evaluation: Importance of Predictors

- It's all about F-tests again!
- Use type II sums of squares to evalute unique effect of predictor
- `Anova` function from `car` library

	Sum Sq Df	F	value	Pr(>F)
distance	2834.545	1	22.520914	0.0000082
firesev	1048.972	1	8.334249	0.0049190
hetero	1617.988	1	12.855173	0.0005578
Residuals	10824.200	86	NA	NA

Evaluation: Estimate of Predictors

- It's all about T-tests again!

E	stimate Std	. Error	t value Pr	(>|t|)
(Intercept)	-2.8676426	9.8658932	-0.2906622	0.7720094
distance	0.7090347	0.1494082	4.7456205	0.0000082
firesev	-2.1654652	0.7500979	-2.8869099	0.0049190
hetero	39.6261529	11.0520528	3.5854111	0.0005578
R^2 = 0.4670	008			

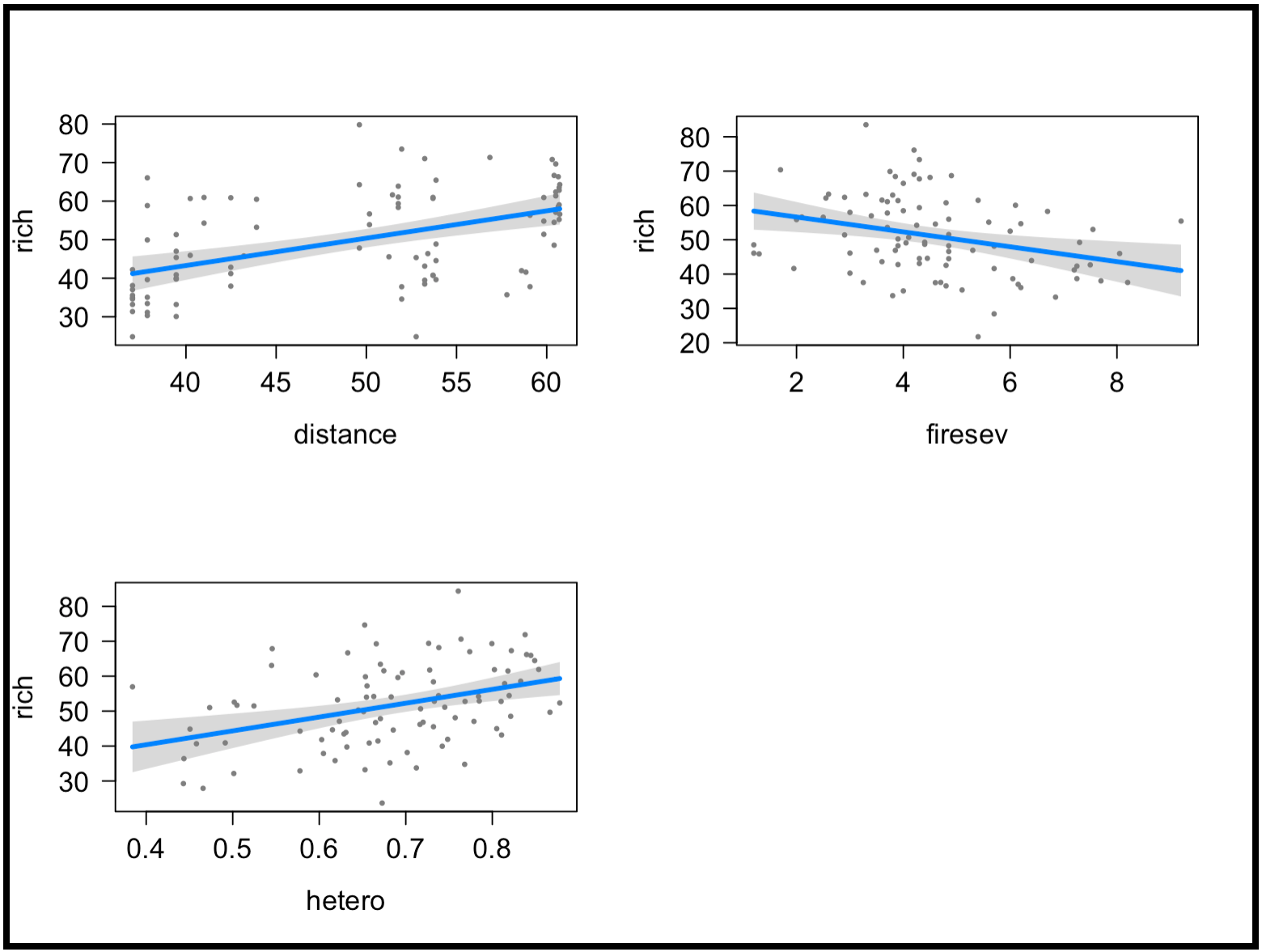
Visualization

- Visualization is difficult with mutiple predictors
  - How do you show effect of multiple continuous variables?
- Strategy 1: Effect of one variable after accounting for effect of others
- Strategy 2: Effect of one variable at different levels of others

Plot Each Predictor at Median of Others

```
library
visreg
```

Plot Each Predictor at Median of Others



What about Interactions?

$$Diversity_i = \beta_0 + \beta_1 Age_i + \beta_2 Severity_i + \beta_3 Age_i * Severity$$

Same Workflow

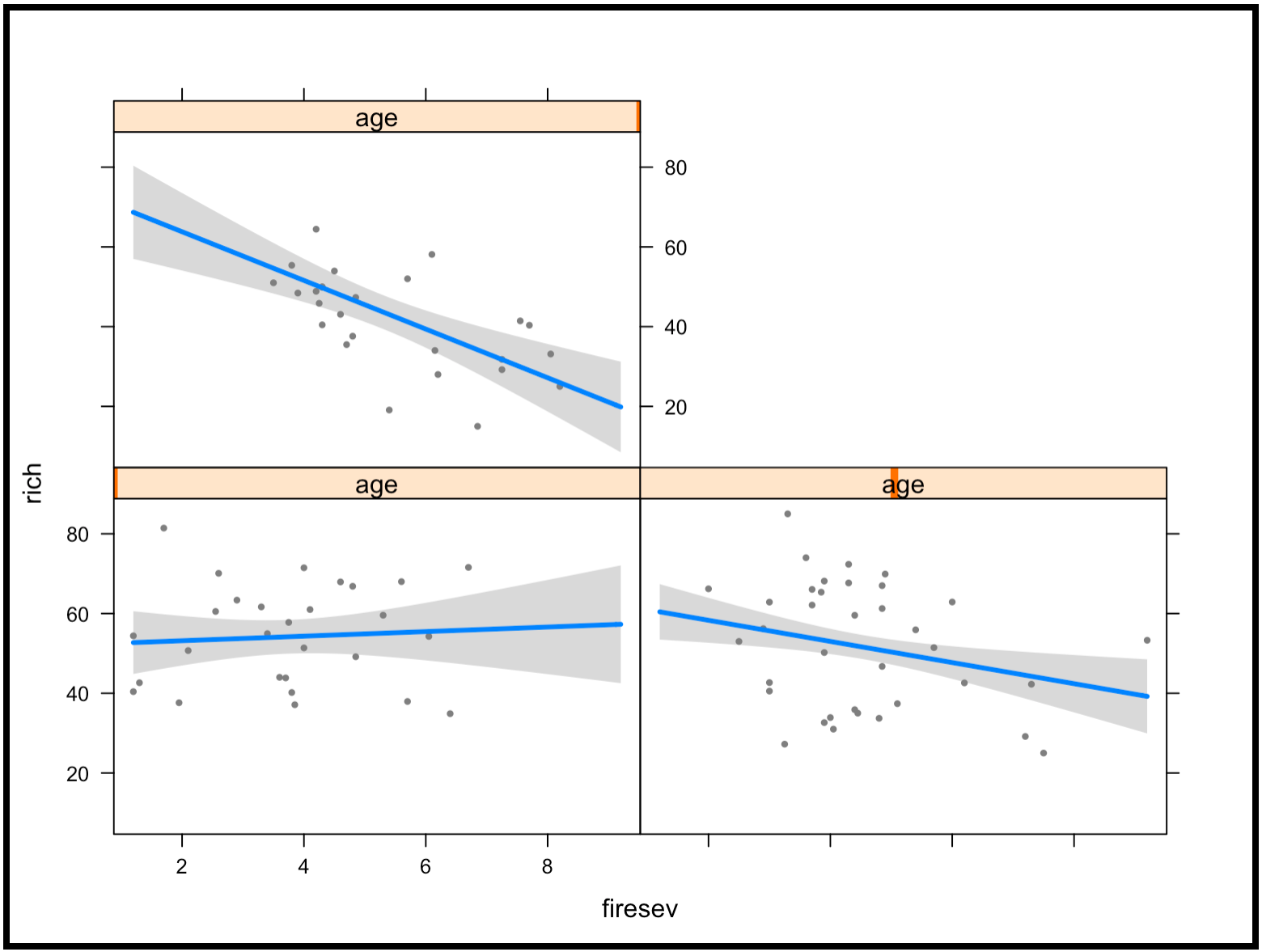
- Evaluate residuals, fitted v. residuals
- Evaluate VIF and correlation of predictors
- F, T-Tests, R^2
- Visualization

F-Tests

	Sum Sq	Df	F value	Pr(>F)
age	466.9263	1	2.701340	0.1039152
firesev	1361.5086	1	7.876824	0.0061927
age:firesev	2228.6867	1	12.893767	0.0005479
Residuals	14865.0939	86	NA	NA

Visualization of One Variable at Levels of Others

```
visreg              "firesev"  by="age"
```


age
age
age
rich
firesev
80
60
40
20
2
4
6
8

Split Data into Quantiles

```
#Make the breaks
                    cut
                        breaks=quantile

                        cut
                        breaks=quantile

#make the plot
ggplot data =          aes x=          y=
  geom_point
  stat_smooth method="lm"
  facet_wrap
```

Split Data into Quantiles