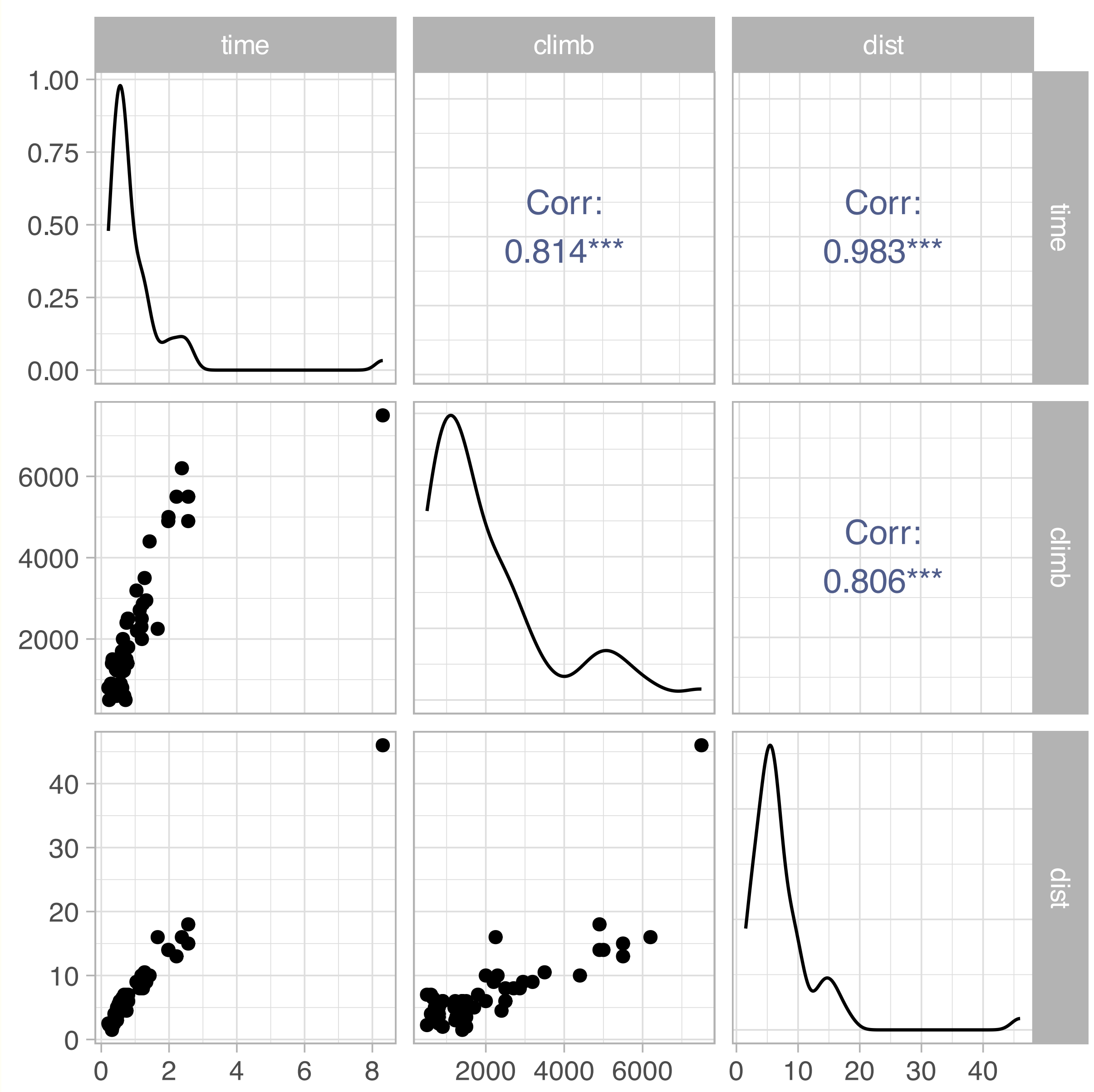
Model Selection

Stat 230: Applied Regression Analysis

Visualizing a fitted model

How can we create a useful 2-dimensional picture of the relationship between Y and x_i, after accounting for the other variables in the model?

The problem with scatterplots



We only see the marginal relationships between pairs of variables, not the relationships after accounting for other variables

Partial residual plots

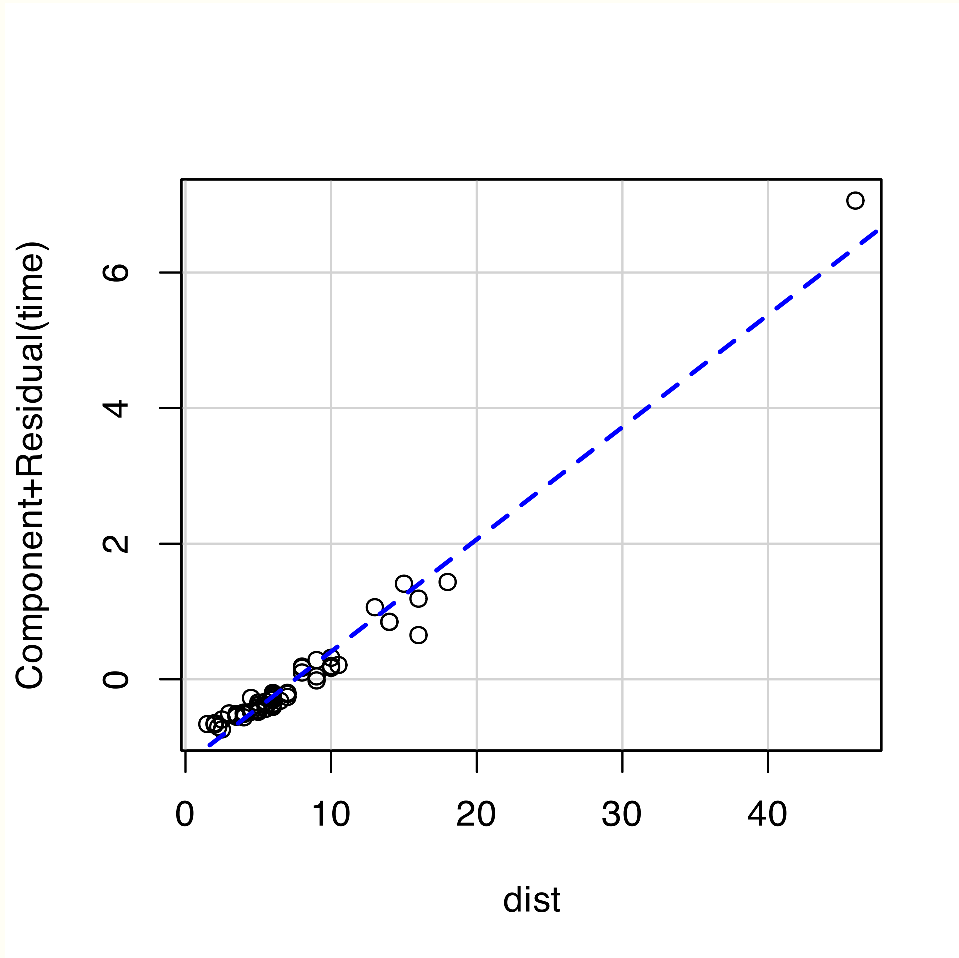
Consider two-predictor model: $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2$

To isolate the relationship between Y and x_2 after accounting for x_1, we can:

1. Fit the MLR model
2. Calculate the residuals from the fitted model: $e_i = y_i - \widehat{y}_i$
3. Add the "contribution" of x_j back into residuals:
   $\text{pres}_{j,i} = e_i + \widehat{\beta}_j x_{j,i}$
4. Plot pres_j against x_j

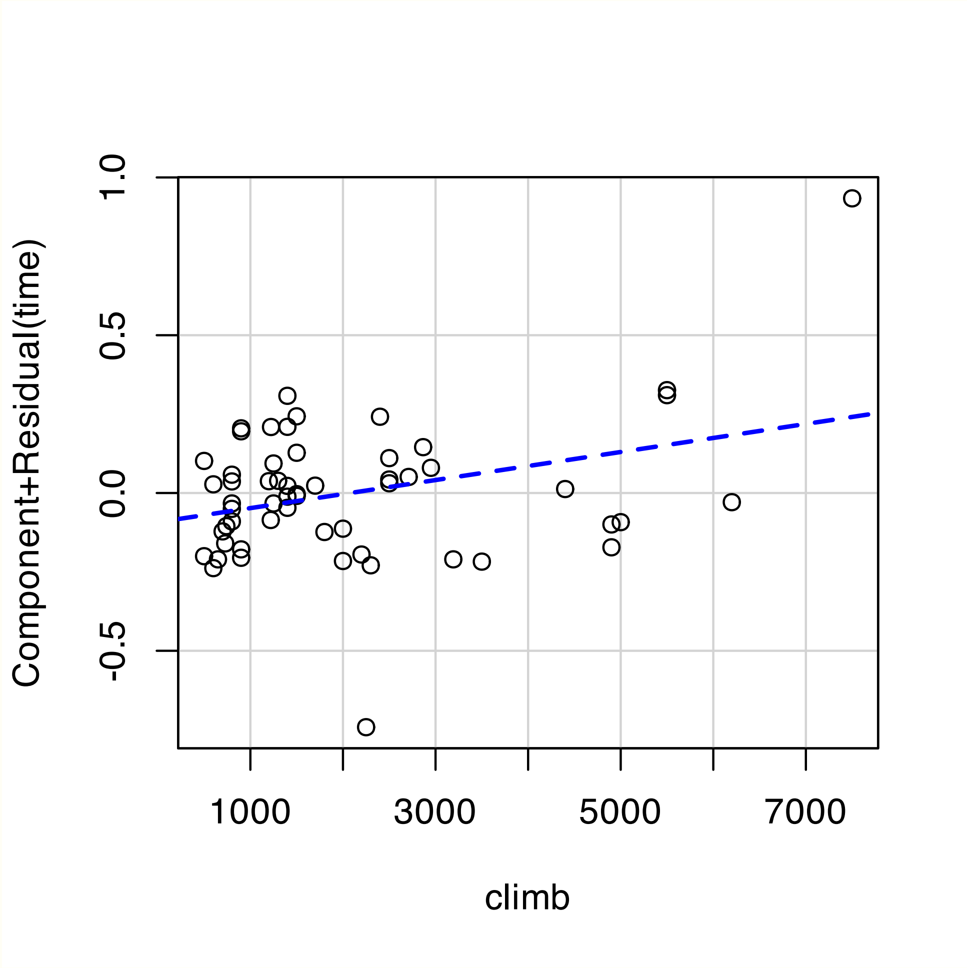
Example

After accounting for climb, what is the relationship between time and distance?



Example

After accounting for distance, what is the relationship between time and climb?



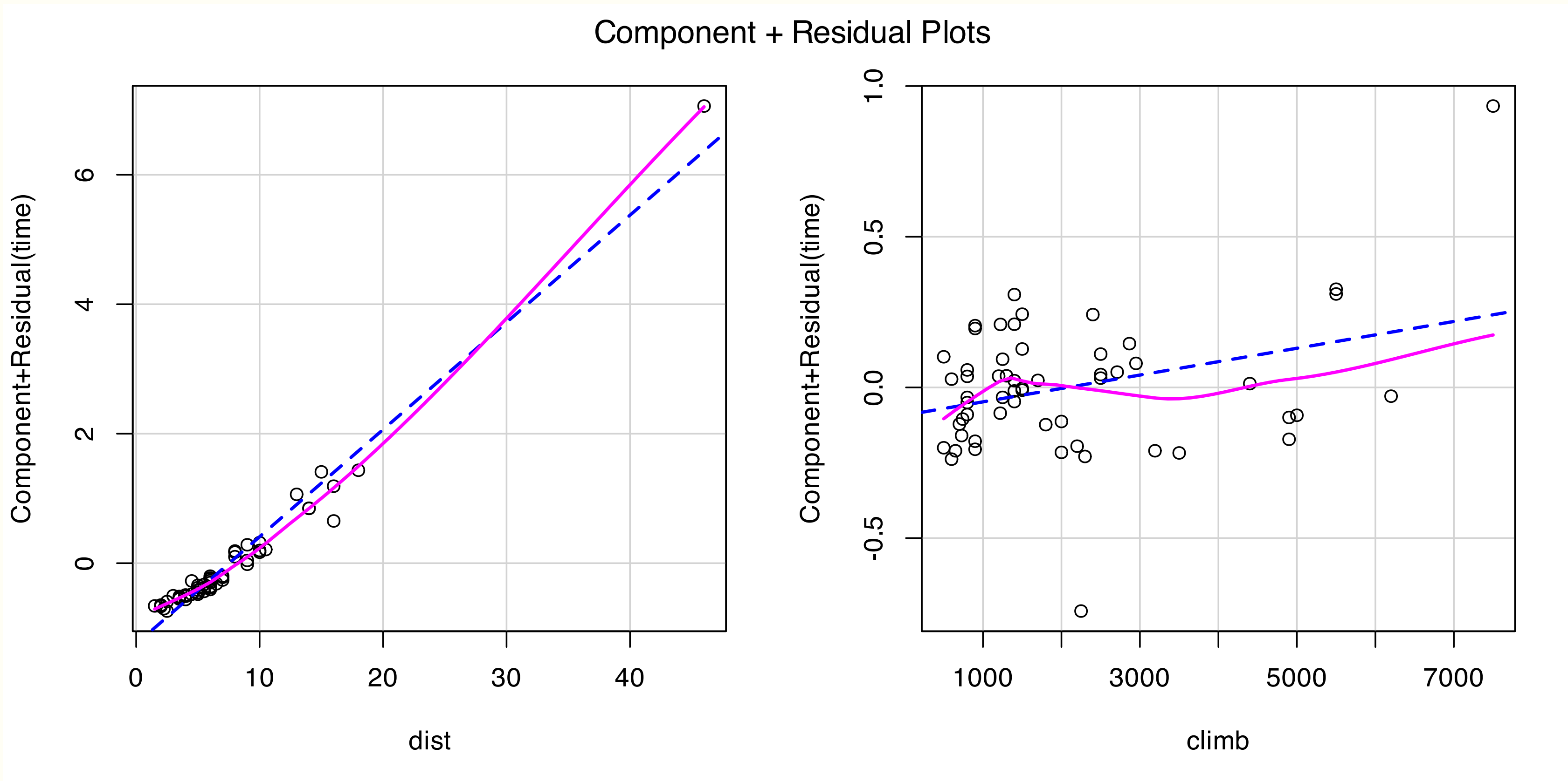
Why is this useful?

- We can see the "effect" of x_j after adjusting for other model terms
- We can see the variation in y that remains after adjusting for other model terms
- We can look for outliers that could be affecting the estimated effect of x_j
- We can see if the effect of x_j is correctly modeled, **non-linearity** and/or non-constant variance suggest we need to correct our model form

Partial residual plots in R

The `car` package calls them **component + residual plots**

```
library(car)
mod <- lm(time ~ dist + climb, data = hills2000)
crPlots(mod, layout = c(1, 2))
```

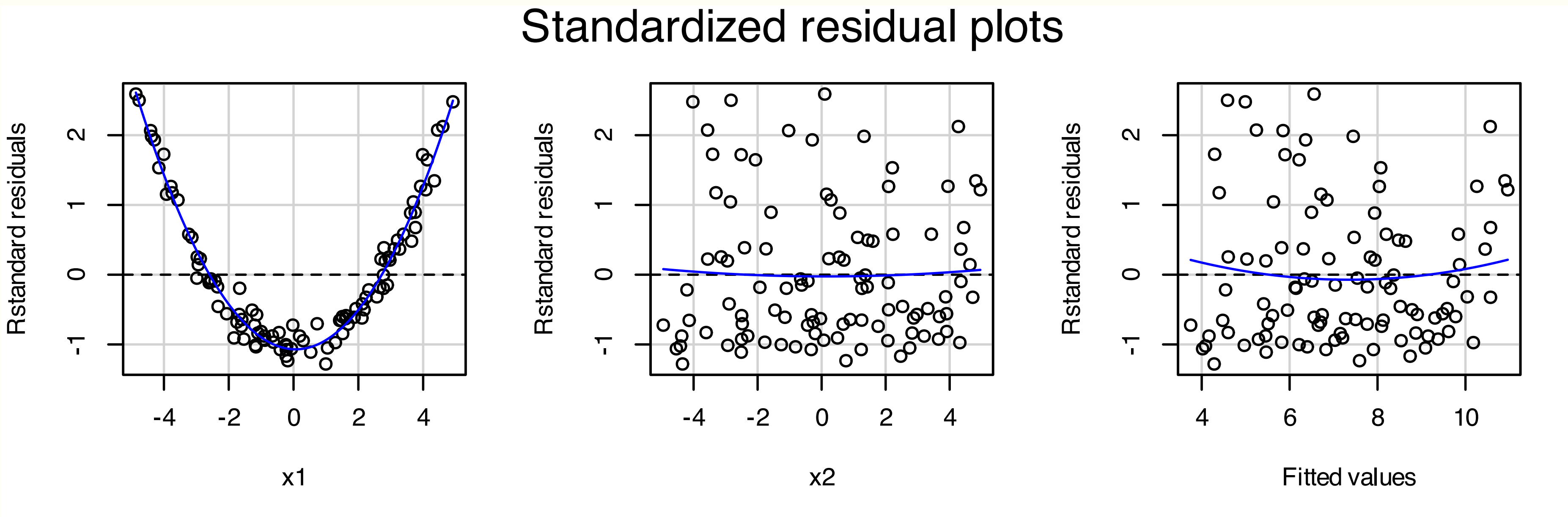


Example 1

True model $\quad y = x_1^2 + 0.5x_2 + \epsilon$

Fitted model $\quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

Both x_1 and x_2 are numeric, roughly between -5 and 5

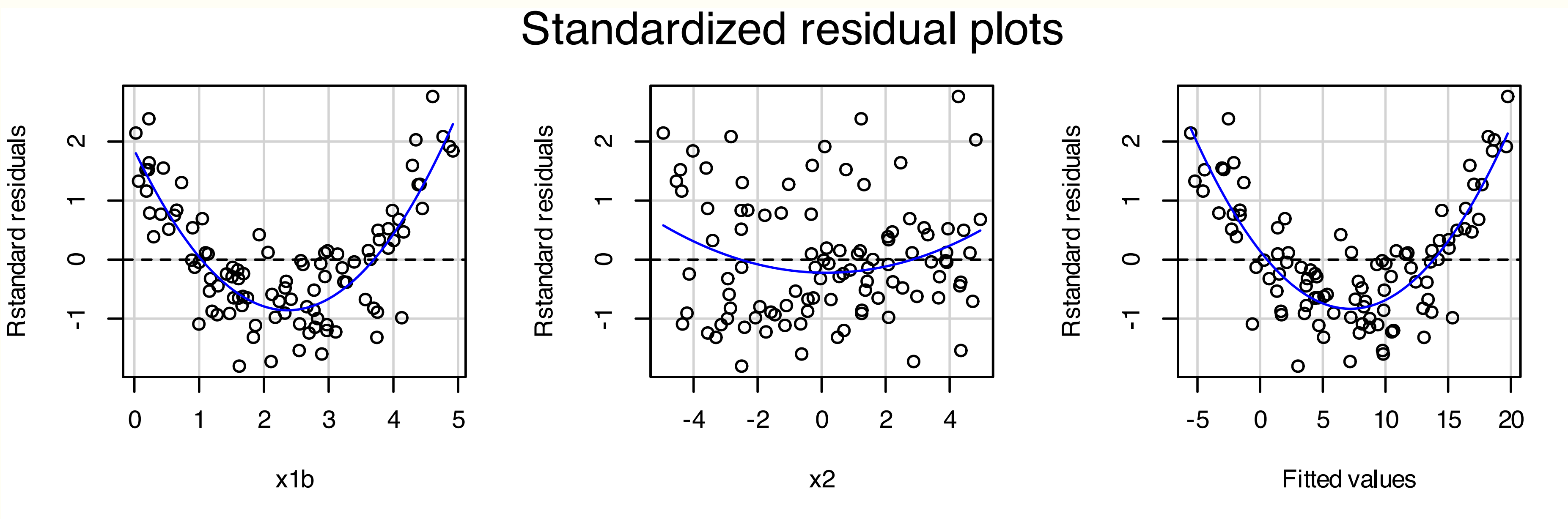


Example 2

True model $y = x_1^2 + 0.5x_2 + \epsilon$

Fitted model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

But now x_1 is strictly positive $\rightarrow$ monotone relationship

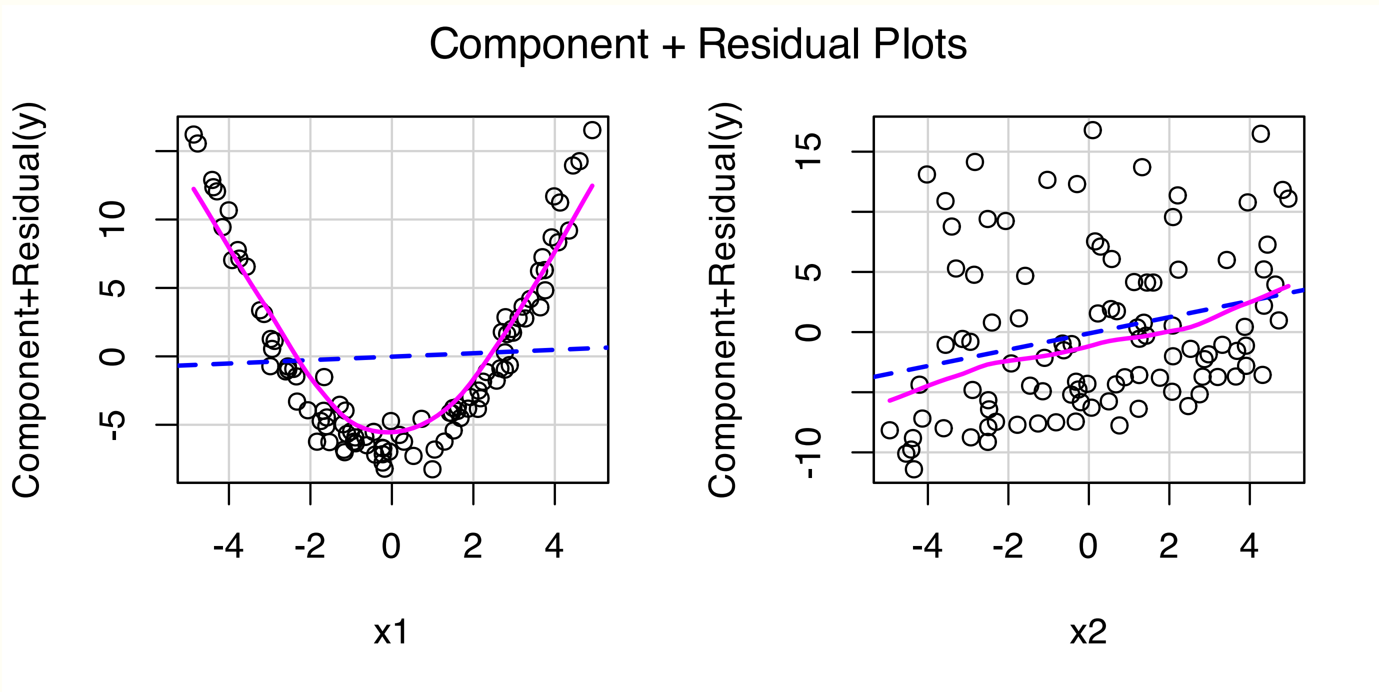


Example 1

True model $y = x_1^2 + 0.5x_2 + \epsilon$

Fitted model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

Both x_1 and x_2 are numeric, roughly between -5 and 5

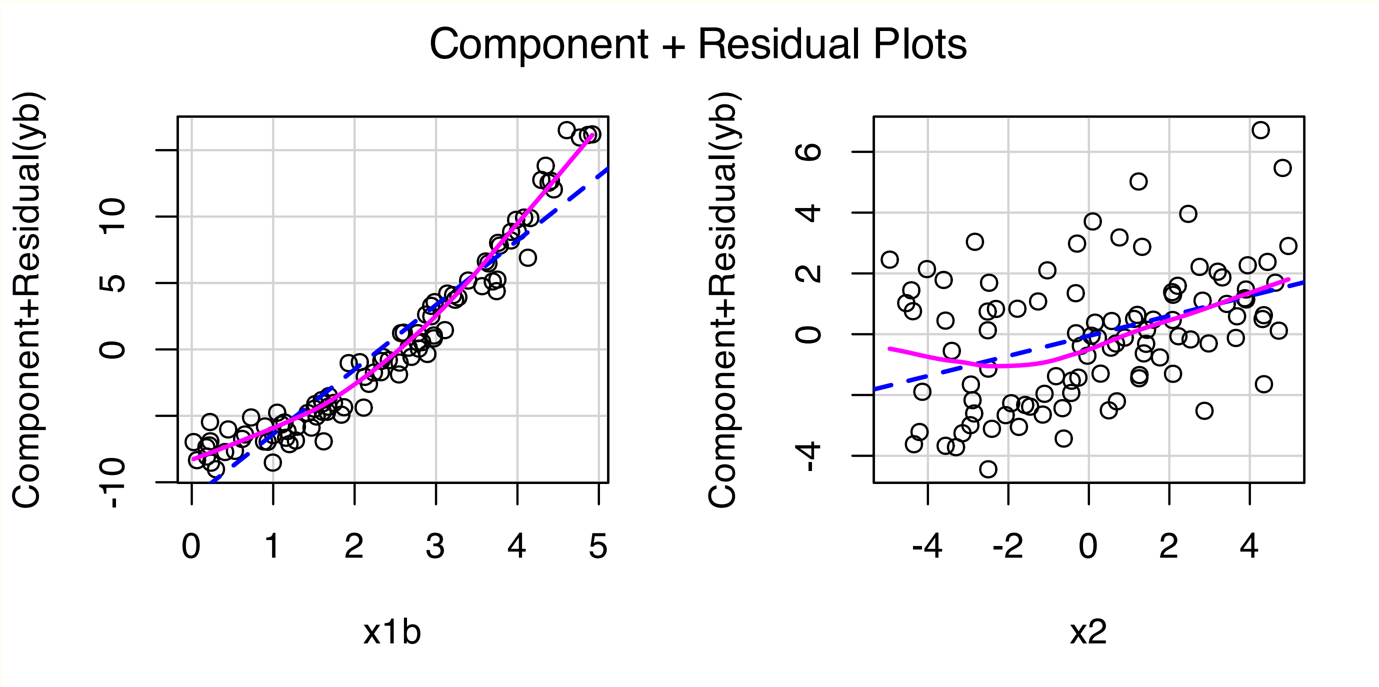


Example 2

True model $y = x_1^2 + 0.5x_2 + \epsilon$

Fitted model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

But now x_1 is strictly positive $\rightarrow$ monotone relationship



Your turn

- Work through example on the handout
- Be ready to share your thoughts (I'm going to cold call)

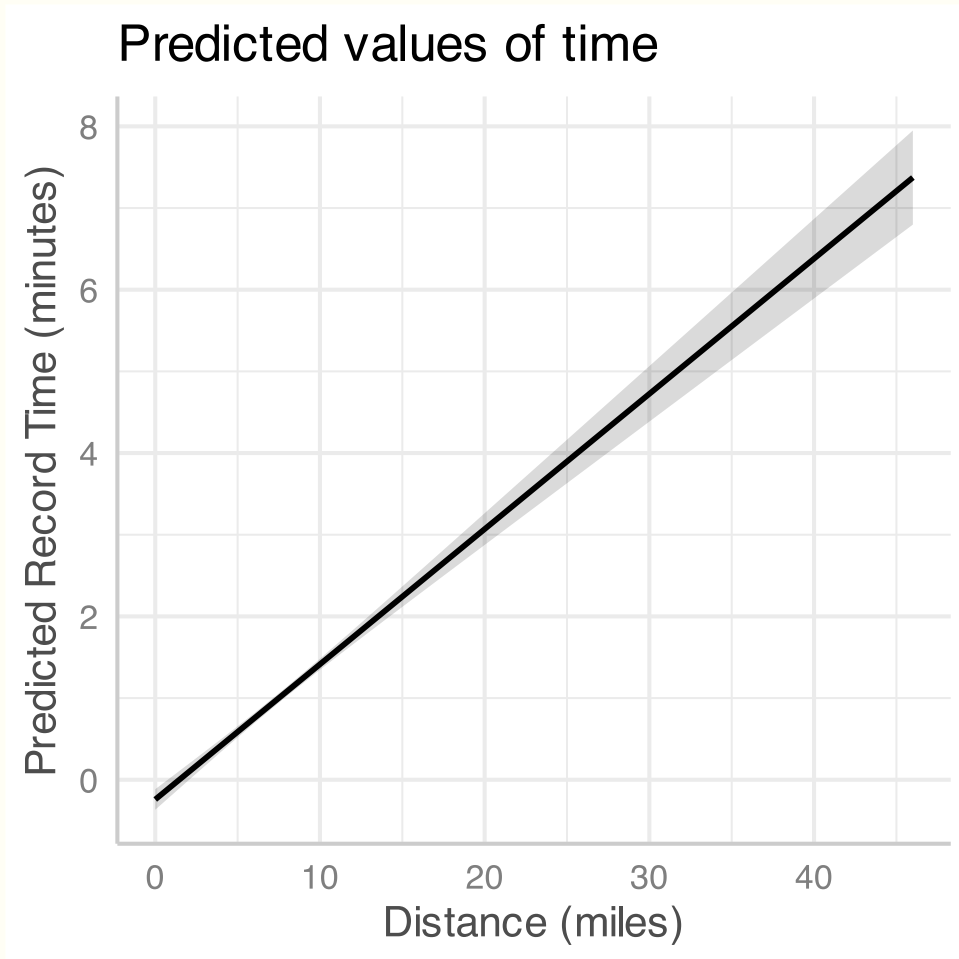
Effects plots

Consider two-predictor model: $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2$

1. Fix x_1 at some value, say $x_1 = c$
2. Calculate $\widehat{y}$ for a range of x_2 values: $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 c + \widehat{\beta}_2 x_2$
3. Plot $\widehat{y}$ against x_2

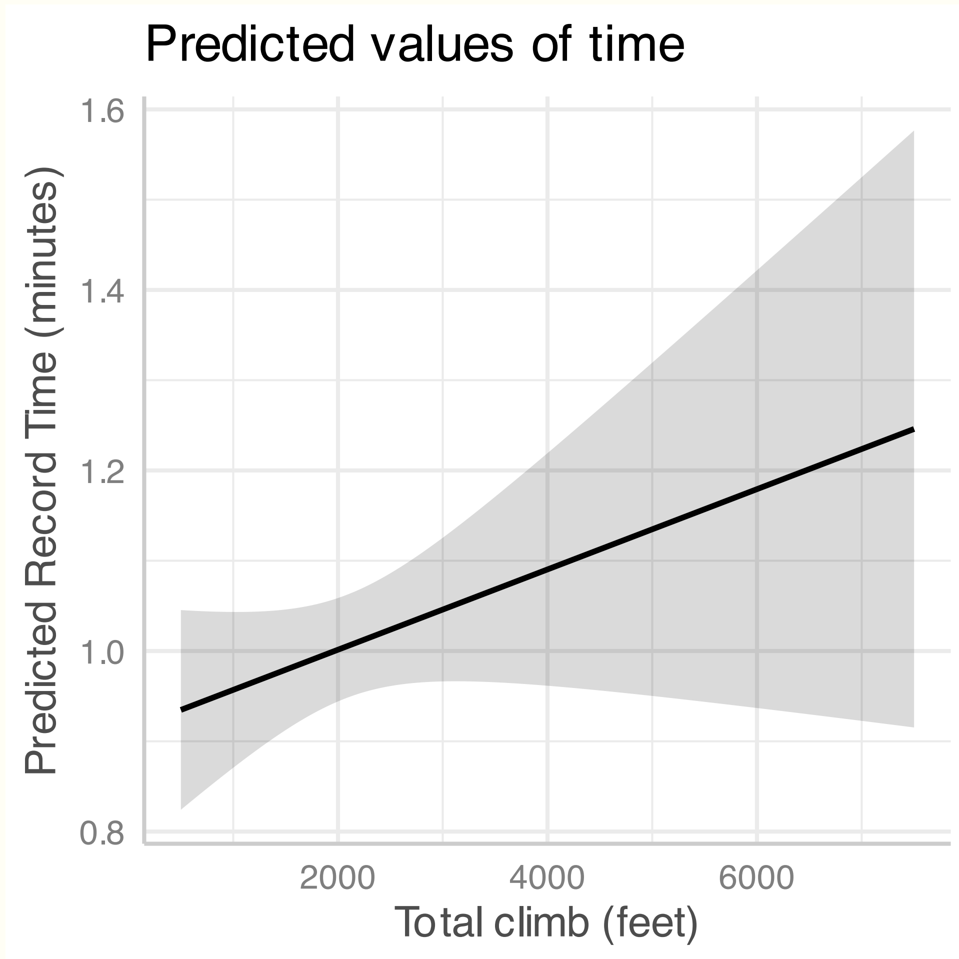
Example

Whats the relationship between record time and distance, holding the total climb constant?



Example

Whats the relationship between record time and the total climb, holding the distance constant?



Effects plots in R

The `ggeffects` package provides a nice way to visualize fitted models

```r
library(ggeffects)
mod <- lm(time ~ dist + climb, data = hills2000)
predict_response(mod, terms = "dist") |>
  plot() +
  labs(
    y = "My y-axis label",
    x = "My x-axis label"
  )
```

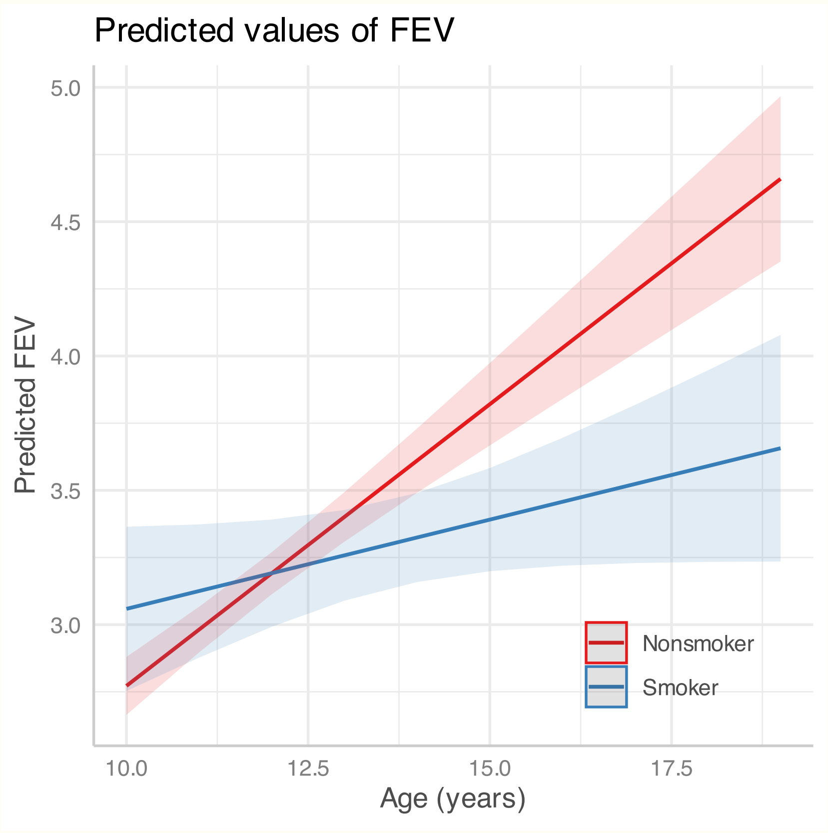
Note

It holds the other predictors at their mean (for numeric) or mode (for categorical)

Plotting interaction models

Recall the FEV model with interaction between age and smoking status:

$$\mu(y|x) = \beta_0 + \beta_1 \texttt{smoker} + \beta_2 \texttt{age} + \beta_3 \texttt{age} \times \texttt{smoker}$$



Effects plots in R

Note

To get multiple fitted lines representing different groups, specify the variable you want to plot and the grouping variable in the `terms` argument

```
fev_lm <- lm(FEV ~ Age * Smoke, data = fev)
predict_response(fev_lm, terms = c("Age", "Smoke")) |>
  plot()
```

Your turn

- Work through example on the handout
- Be ready to share your thoughts (I'm going to cold call)

Model building

1. Define the goal

Before you start building a model you need to identify why you are building the model.

- Exploring associations
- Testing a theoretical relationship
- Controlling for confounders
- Prediction

2. Choose an initial pool of predictors

- Theory might dictate some/all variables
- Designed experiment might dictate some/all variables
- In other situations
  - Examine variables one at a time – beware of skew, note outliers
  - Examine pairwise correlations/scatterplot matrix – note potential predictors, multicollinearity

3. Fit a full regression model

Fit a “full” initial regression model where you include all of the potential variables

4. Question your full model

- Check the full model for violations to the conditions, fix as needed.
- Order I check/fix:

1. Linearity
2. Heteroscedasticity
3. Normality
4. Outliers and influential points

5. Examine if any variables can be dropped/added

- There may be "insignificant" predictor variables that you can consider dropping
- You could use t-tests or extra-sums-of-squares F-tests to guide these decisions
- You could use model selection criteria (AIC, BIC, adjusted R^2) to guide these decisions
- Sometimes you discover reasons to add variables (e.g., remedy model deficiencies, discovery of interactions)

6. Iterate through steps 4 and 5

- Modeling is an iterative process, unlikely to find the "best" model on the first try
- Each time you change the model, you need to re-check/fix the model conditions

7. Do a final model check

- Are the conditions are satisfied?
- Outliers and influential points?
- Multicollinearity?

8. Proceed with your analysis

- Interpret coefficients
- Test hypotheses
- Make predictions

Confirming a theory

When you want to confirm a theory, only include "extra" predictors in the model building process.

Add the variables that are "predetermined" by the theory back into the model at the end of the model building process.

SAT data

Data for the 50 states

Variable	Description
`sat`	average of combined verbal and math SAT
`takers`	percentage of eligible seniors who took exam
`income`	median income of families of test-takers
`years`	mean number of years of schooling
`public`	percentage of test-takers attending public school
`expend`	total state expenditure on secondary schools (in hundreds of dollars per student)
`rank`	median percentile rank of test-takers in their high school classes

Working for the legislature

What is the impact of state expenditures on SAT scores after accounting for other factors?

Strategy: First, choose controls, then add expenditures

term	estimate	std.error	statistic	p.value
(Intercept)	144.5300	297.1628	0.4864	0.6291
Rank	4.4498	2.7466	1.6201	0.1124
Income	0.2054	0.1162	1.7672	0.0841
Years	24.8751	6.4223	3.8732	0.0004
log(Takers)	-26.0915	17.0035	-1.5345	0.1321
Public	0.6962	0.5513	1.2630	0.2132

- `Rank`, `log(Takers)`, and `Public` may not be significant - but can we trust these results?

Our model is overspecified!

`Rank` and `log(Takers)` are highly correlated!

```
vif(political_lm)
```

```
     Rank    Income     Years log(Takers)    Public 
   21.068     1.692     1.326      22.175     1.928 
```

Let's try dropping `Rank` and refitting the model…

Model B

It looks like `Public` can also be removed as it doesn't explain a substantial proportion of the variability in `SAT` scores

term	estimate	std.error	statistic	p.value
(Intercept)	-268.779	127.4007	-2.1097	0.0405
Rank	8.509	0.7496	11.3522	0.0000
Income	0.230	0.1168	1.9689	0.0551
Years	27.564	6.2710	4.3954	0.0001
Public	0.267	0.4821	0.5539	0.5824

Now, add expenditure

Examine the significance of expenditure after controlling for rank, income and years.

term	estimate	std.error	statistic	p.value
(Intercept)	-285.4613	98.9617	-2.885	0.0060
Rank	9.3411	0.7393	12.636	0.0000
Income	0.1199	0.1078	1.112	0.2719
Years	25.5321	5.3994	4.729	0.0000
Expend	1.6162	0.6706	2.410	0.0201