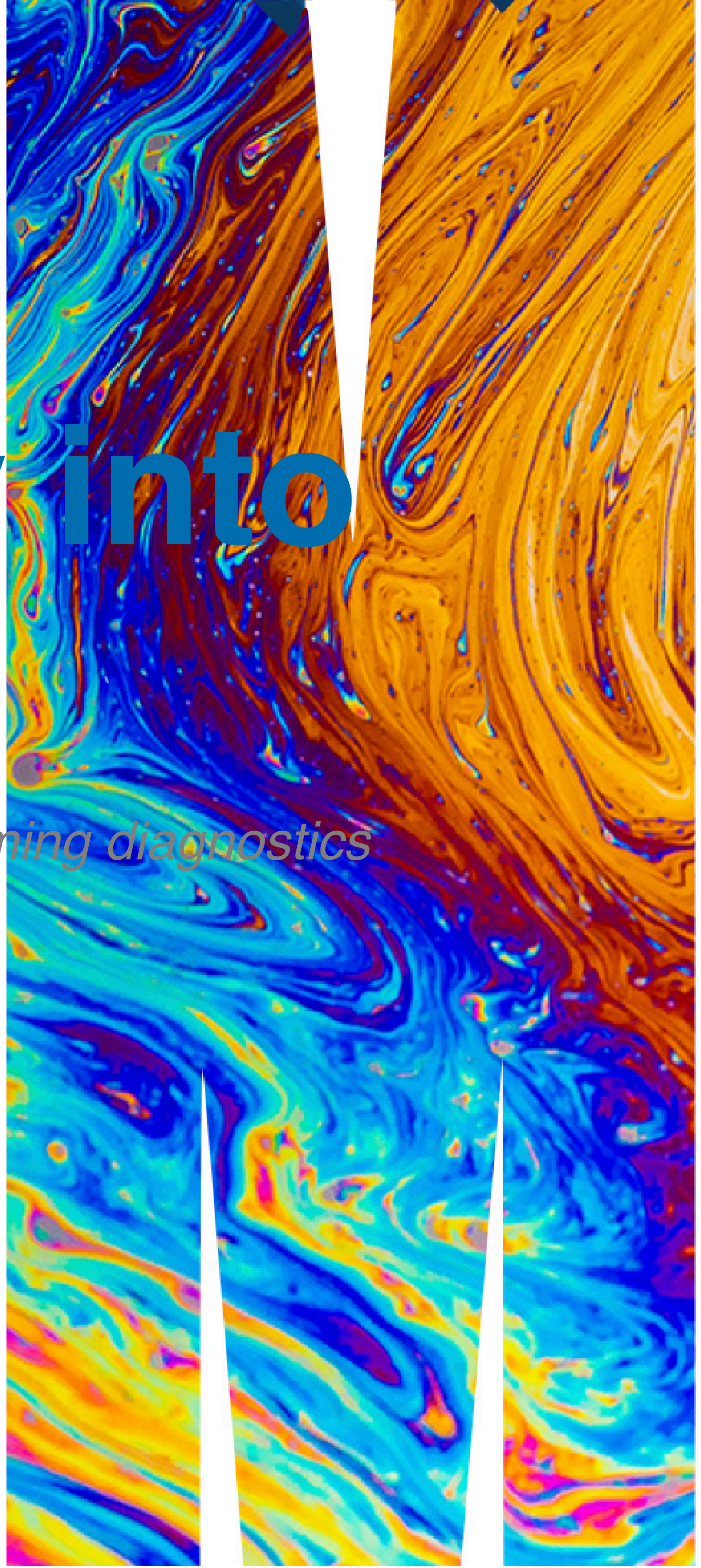
MONASH University

ETC5521: Diving Deeply into Data Exploration

Sculpting data using models, checking assumptions, co-dependency and performing diagnostics

Professor Di Cook

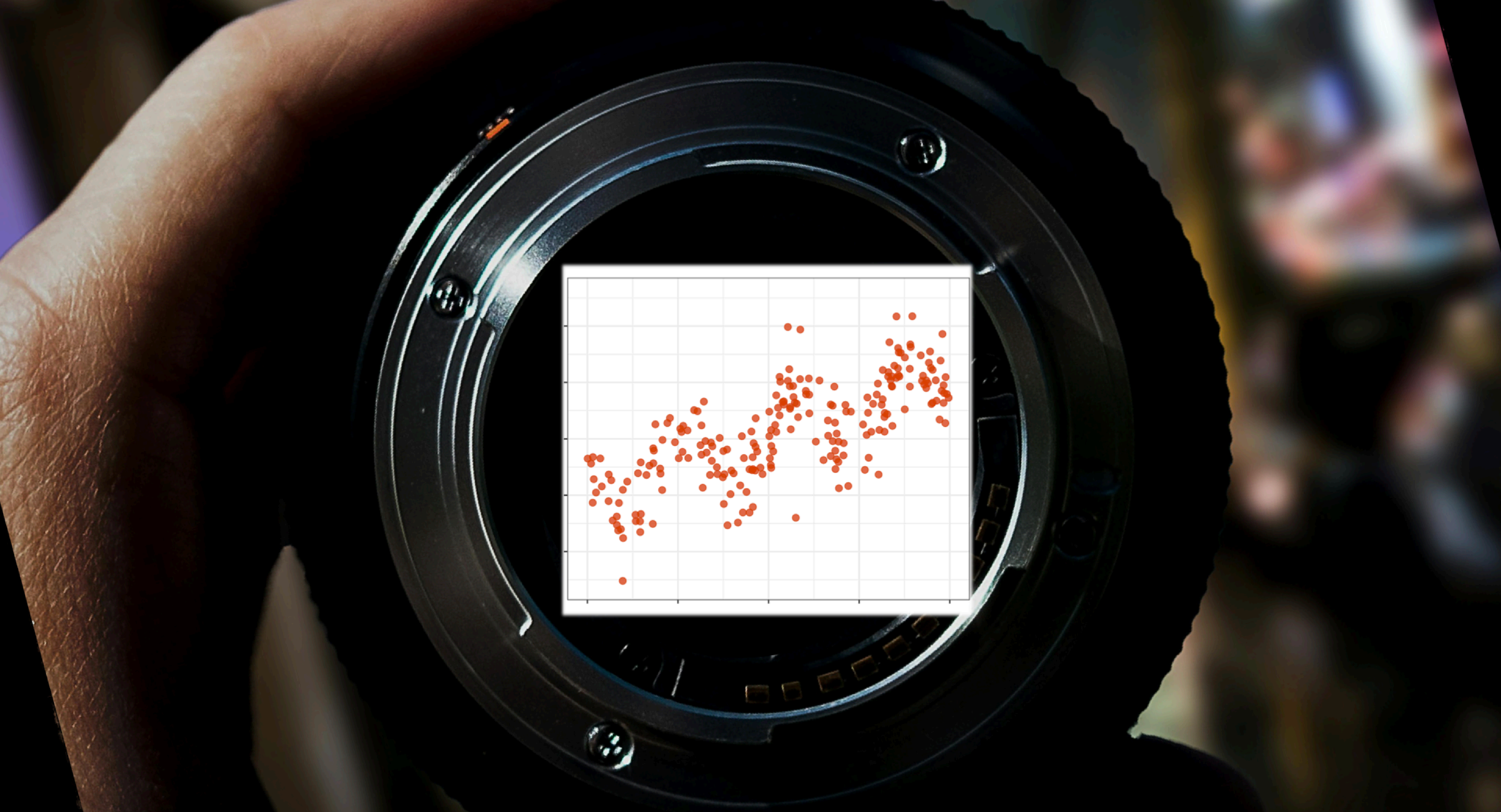
Department of Econometrics and Business Statistics

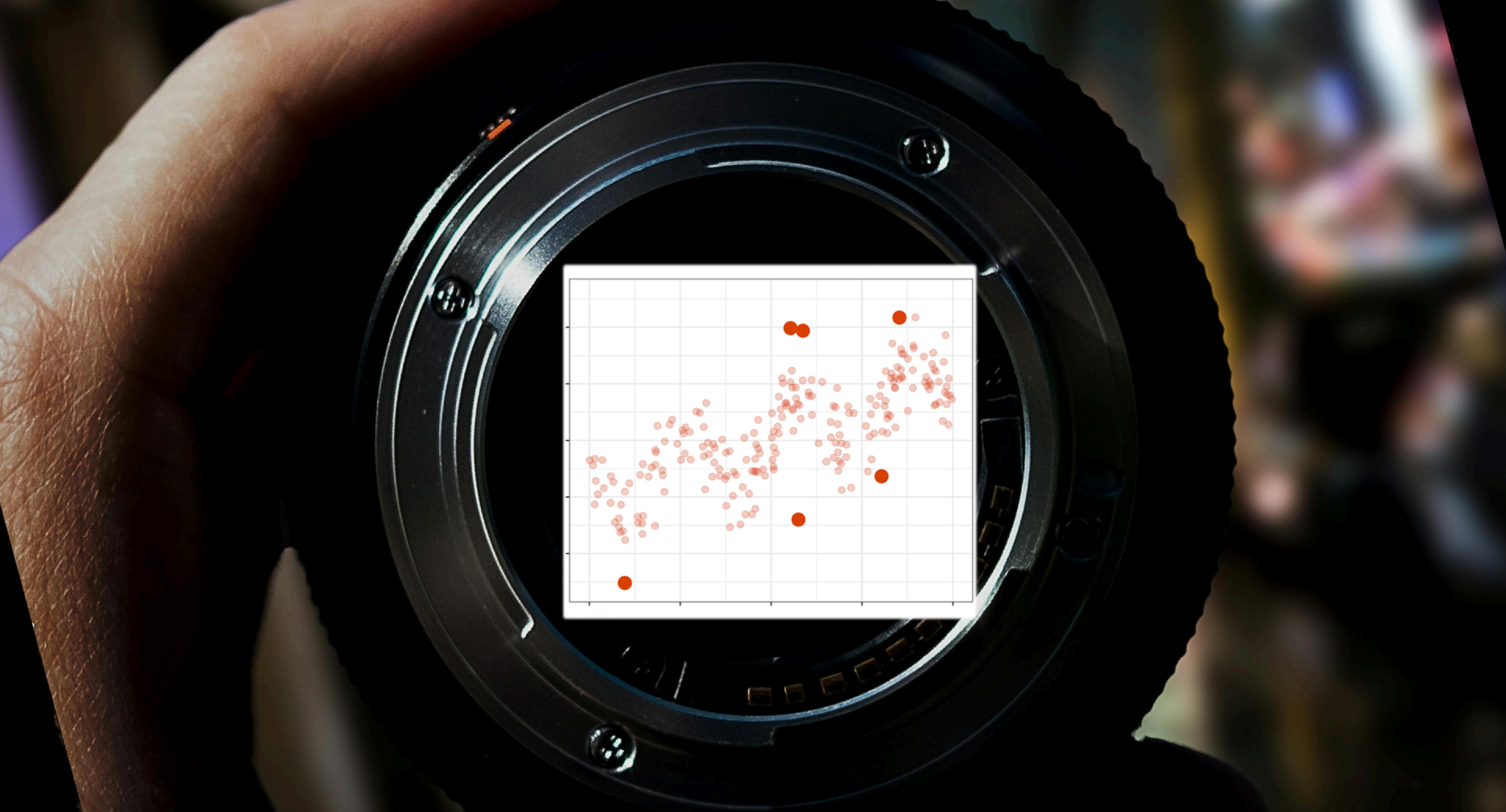


Outline

- Different types of model fitting
- Decomposing data from model
  - fitted
  - residual
- Diagnostic calculations
  - anomalies
  - leverage
  - influence

Models can be used to re-focus the view of data





Different types of model fitting

The basic form for fitting a model with data (response Y and predictors X) is:

$$Y = f(X) + \varepsilon$$

and X could be include multiple variables, $X = (X_1, X_2, \dots, X_p)$ where p is the number of variables. We have a sample of n observations,

$y_i, x_{i1}, \dots x_{ip}, \quad i = 1, \dots, n.$

- In a parametric model, the form of f is specified, e.g. $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$, and one would estimate the parameters $\beta_0, \beta_1, \beta_2, \beta_3$.
  - Frequentist fitting assumes that parameters are fixed values.
  - In a Bayesian framework, the parameters are assumed to have a distribution, e.g. Gaussian.
- In a non-parametric model, the form of f is NOT specified but fitted from the data. May not have a specific functional form, and needs more data, typically. Imposes less assumptions. Can be done in a Bayesian framework.
- Different types of variables can change the model specification, e.g. binary or categorical Y, or temporal or spatial context.
- Different model products, e.g. fitted values or residuals, after the fit change the lens with which we view the data.

Parametric regression

Specification

Specify the

- functional form, e.g. function form is has linear and quadratic terms

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2$$

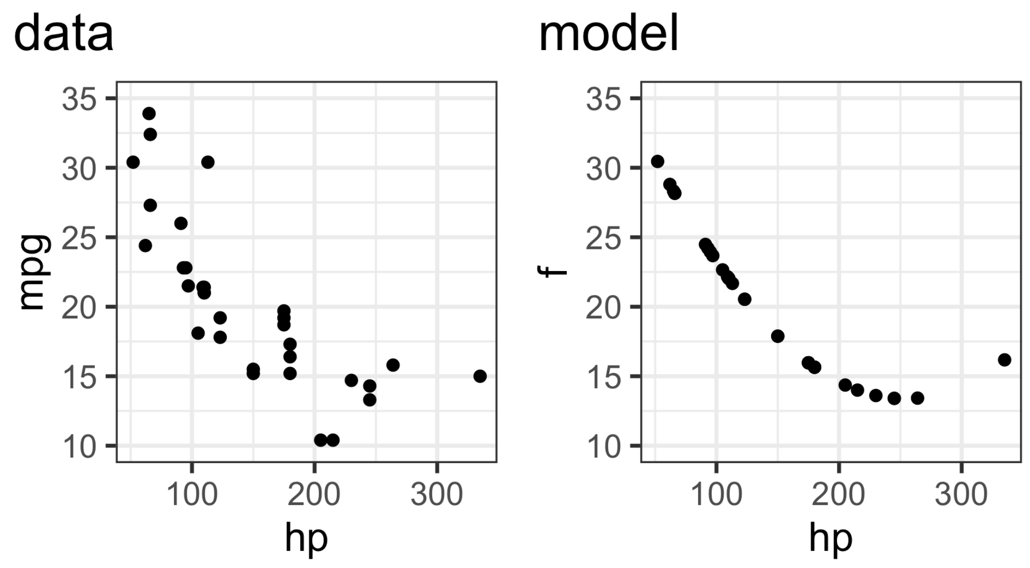
- distribution of errors, e.g.

$$\varepsilon \sim N(0, \sigma^2)$$

Fitting results in:

- fitted values, $\hat{y}$ (sharpening)
- residuals, $e = y - \hat{y}$ (what did we miss)

▶ Code



▶ Code

```
# A tibble: 3 × 5
  term          estimate std.error statistic  p.value
  <chr>            <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)       20.1     0.544     36.9  6.15e-26
2 poly(hp, 2)1     -26.0     3.08      -8.46 2.51e- 9
3 poly(hp, 2)2      13.2     3.08       4.27 1.89e- 4
```

▶ Code

```
# A tibble: 1 × 12
  r.squared adj.r.squared sigma statistic       p.value    df
      <dbl>         <dbl> <dbl>     <dbl>         <dbl> <dbl>
1     0.756         0.739  3.08      45.0       1.30e-9     2
# ℹ 6 more variables: logLik <dbl>, AIC <dbl>, BIC <dbl>,
#   deviance <dbl>, df.residual <int>, nobs <int>
```

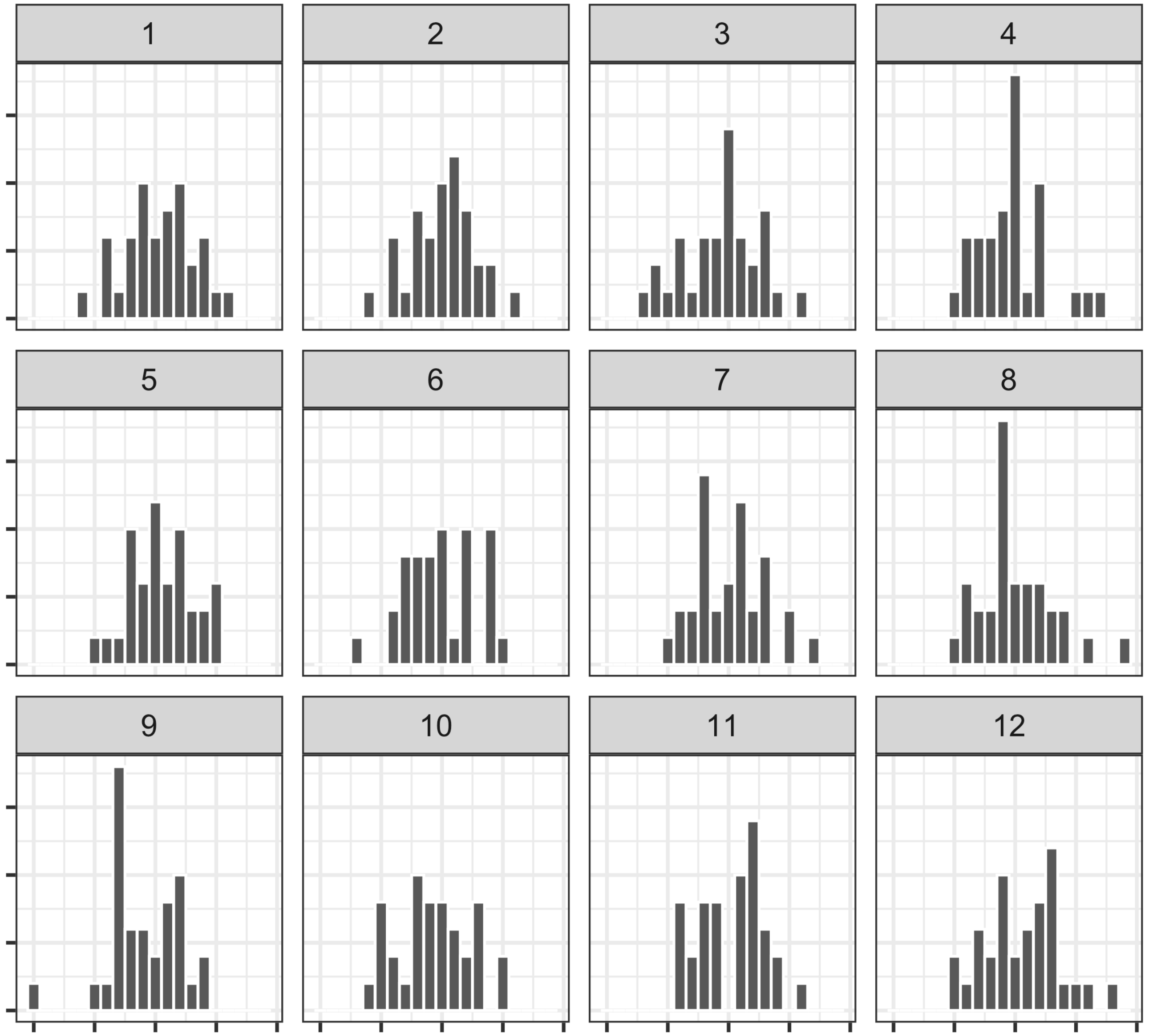
Diagnostics (1/3)

Residuals, $e = y - \hat{y}$ (*what doesn't the fitted model see?*)

- Should be consistent with a sample from the specified error model
- Should have no relationship with the response variable

Lineup | Normal? | Lineup | Relationship?

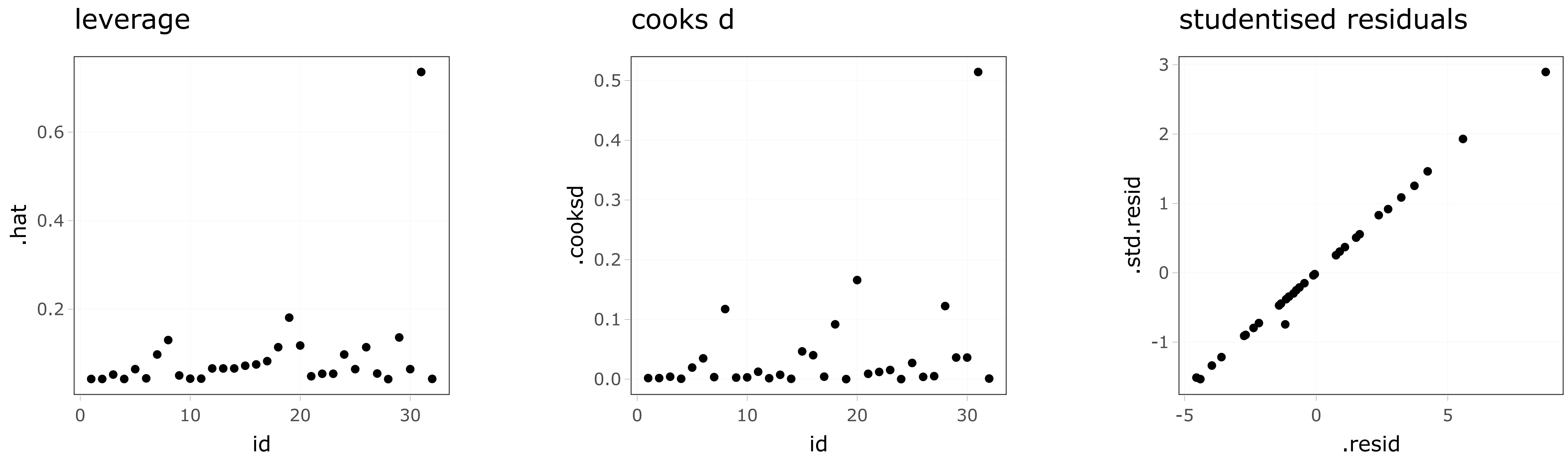
▶ Code



Diagnostics (2/3)

Diagnostics (3/3)

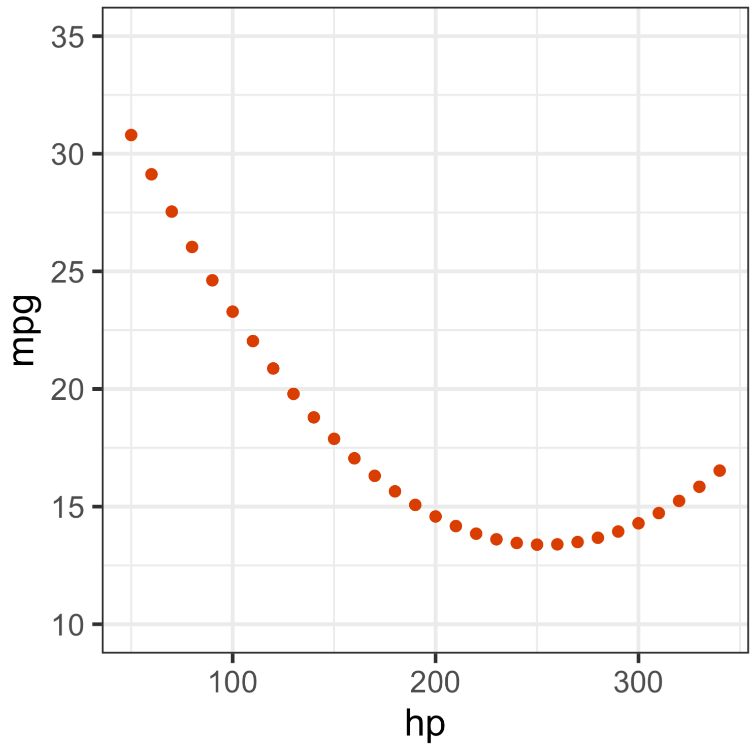
▶ Code



Simulation

Generate response values for un-collected predictor values

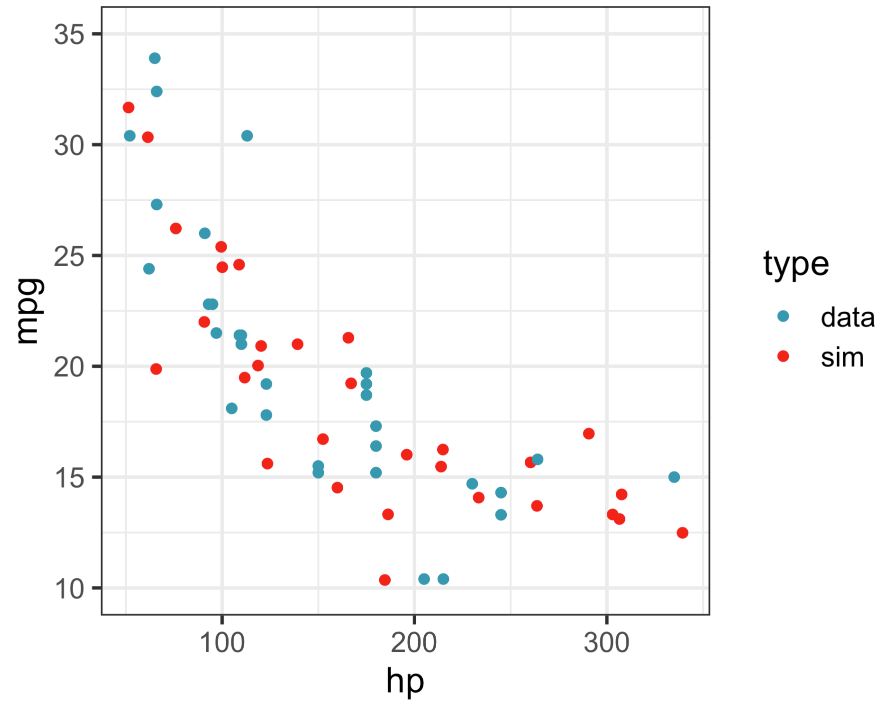
```
mt_full_fit <- tibble(hp = seq(50, 340, 10))
mt_full_fit <- mt_full_fit |>
                 mutate(mpg = predict(mtcars_fi
```



Simulate new samples

1 2 3

▶ Code



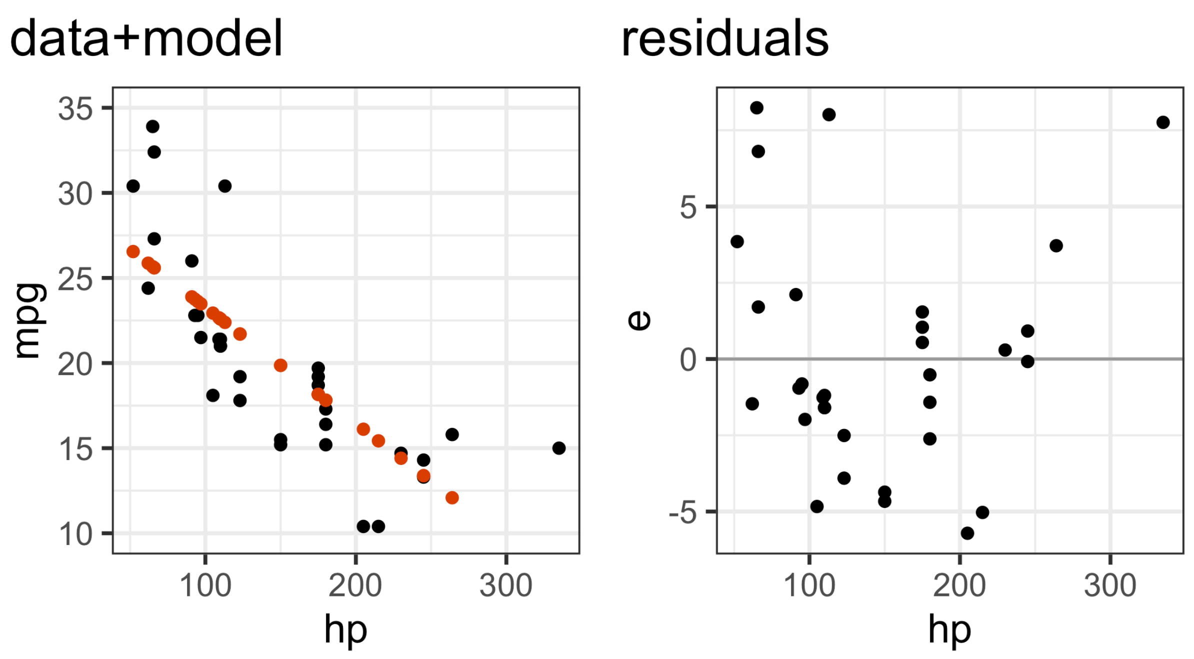
What can go wrong with parametric model?

Wrong specification

Specify function form is has only linear term

$$f(X) = \beta_0 + \beta_1 X$$

- Code



Polynomial

```
# A tibble: 3 × 5
  term          estimate std.error statistic  p.value
  <chr>            <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)       20.1     0.544     36.9  6.15e-26
2 poly(hp, 2)1     -26.0     3.08      -8.46 2.51e- 9
3 poly(hp, 2)2      13.2     3.08       4.27 1.89e- 4
```

```
# A tibble: 1 × 12
  r.squared adj.r.squared sigma statistic      p.value    df
      <dbl>         <dbl> <dbl>     <dbl>        <dbl> <dbl>
1     0.756         0.739  3.08      45.0      1.30e-9     2
# ℹ 6 more variables: logLik <dbl>, AIC <dbl>, BIC <dbl>,
#   deviance <dbl>, df.residual <int>, nobs <int>
```

Linear

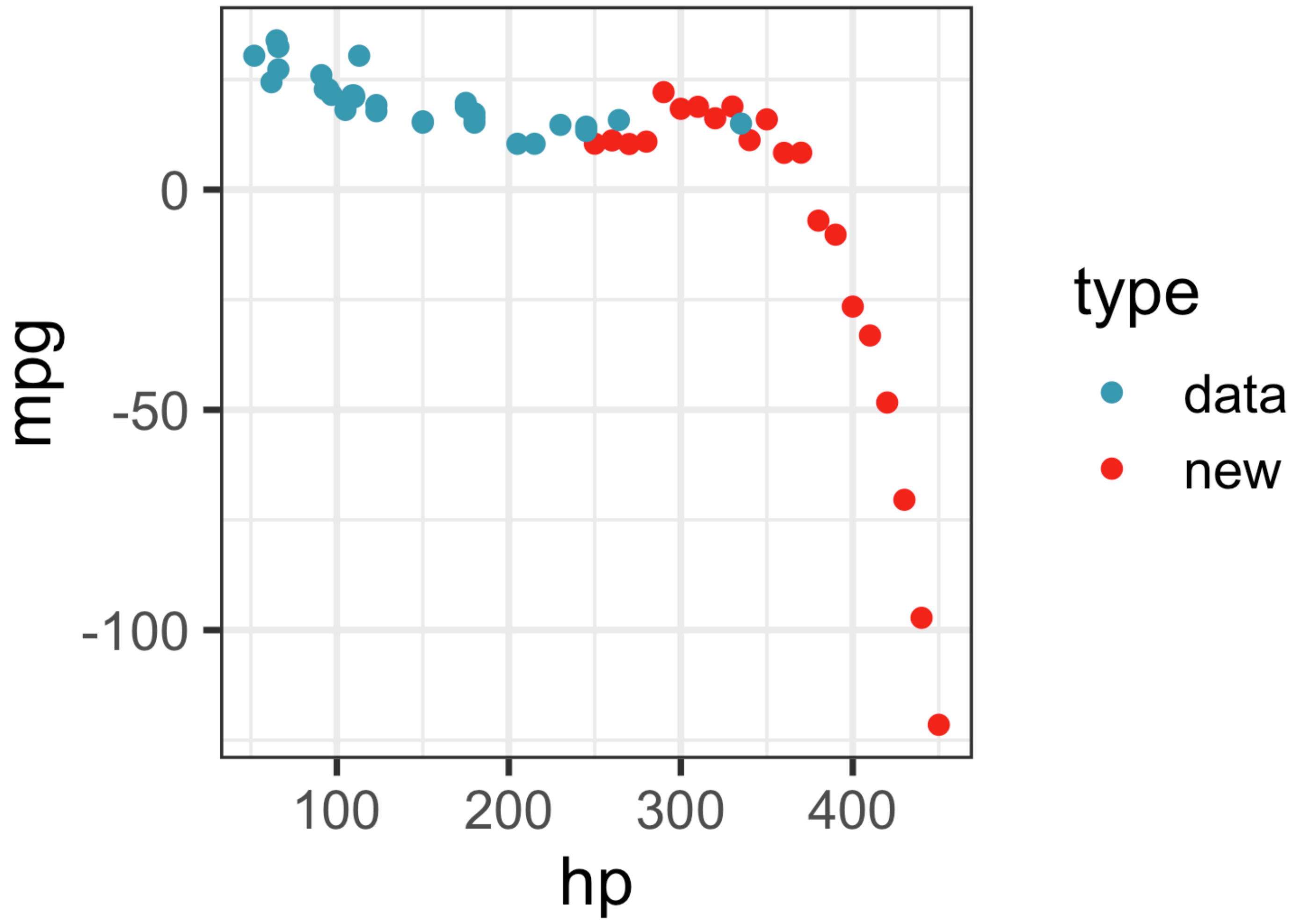
```
# A tibble: 2 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  30.1       1.63       18.4  6.64e-18
2 hp           -0.0682    0.0101     -6.74 1.79e- 7
```

```
# A tibble: 1 × 12
  r.squared adj.r.squared sigma statistic     p.value    df
      <dbl>         <dbl> <dbl>     <dbl>       <dbl> <dbl>
1     0.602         0.589  3.86      45.5 0.000000179     1
# ℹ 6 more variables: logLik <dbl>, AIC <dbl>, BIC <dbl>,
#   deviance <dbl>, df.residual <int>, nobs <int>
```

Extrapolating

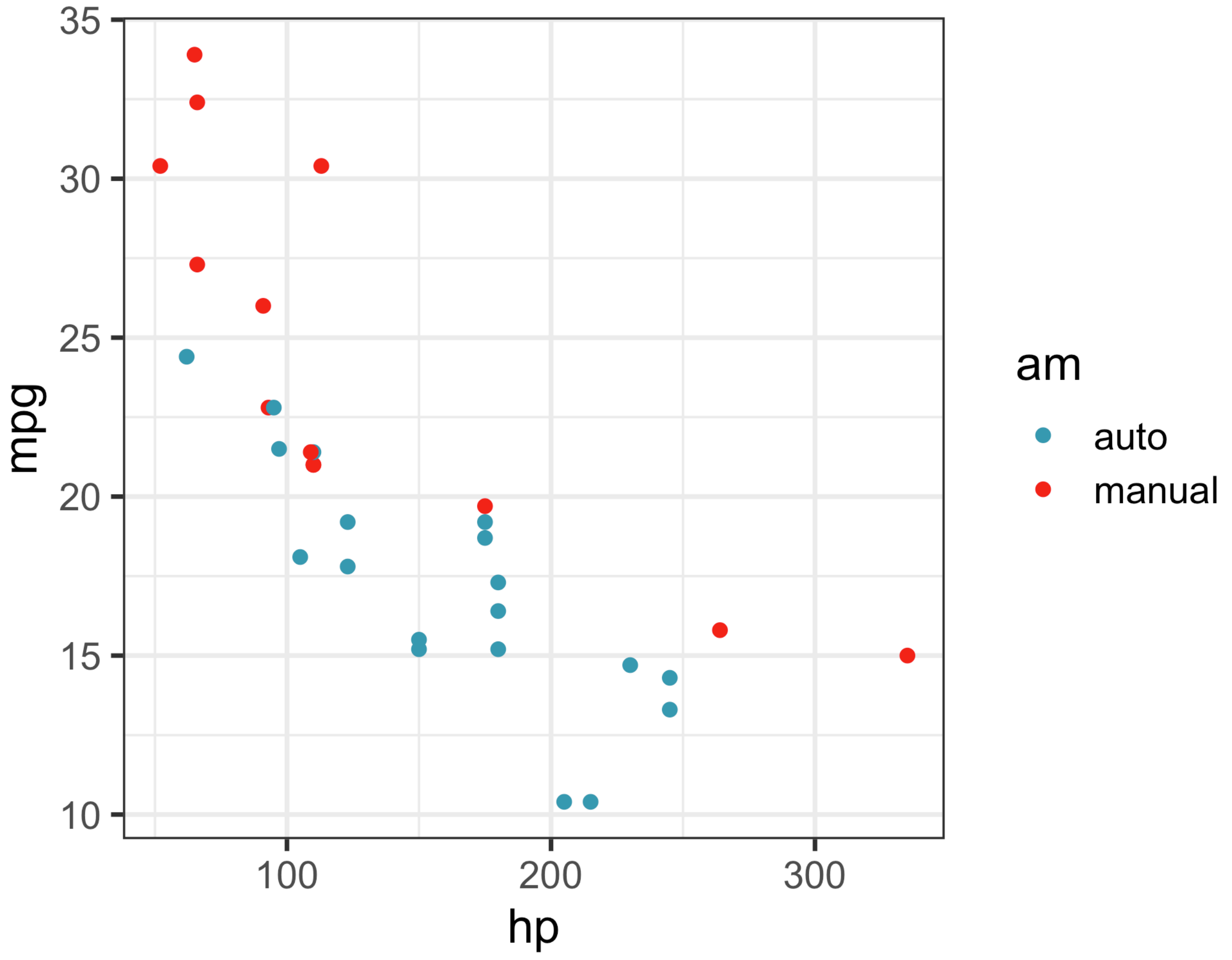
Generate response values for un-collected predictor values OUTSIDE of domain of collected data, can produce HALLUCINATIONS.

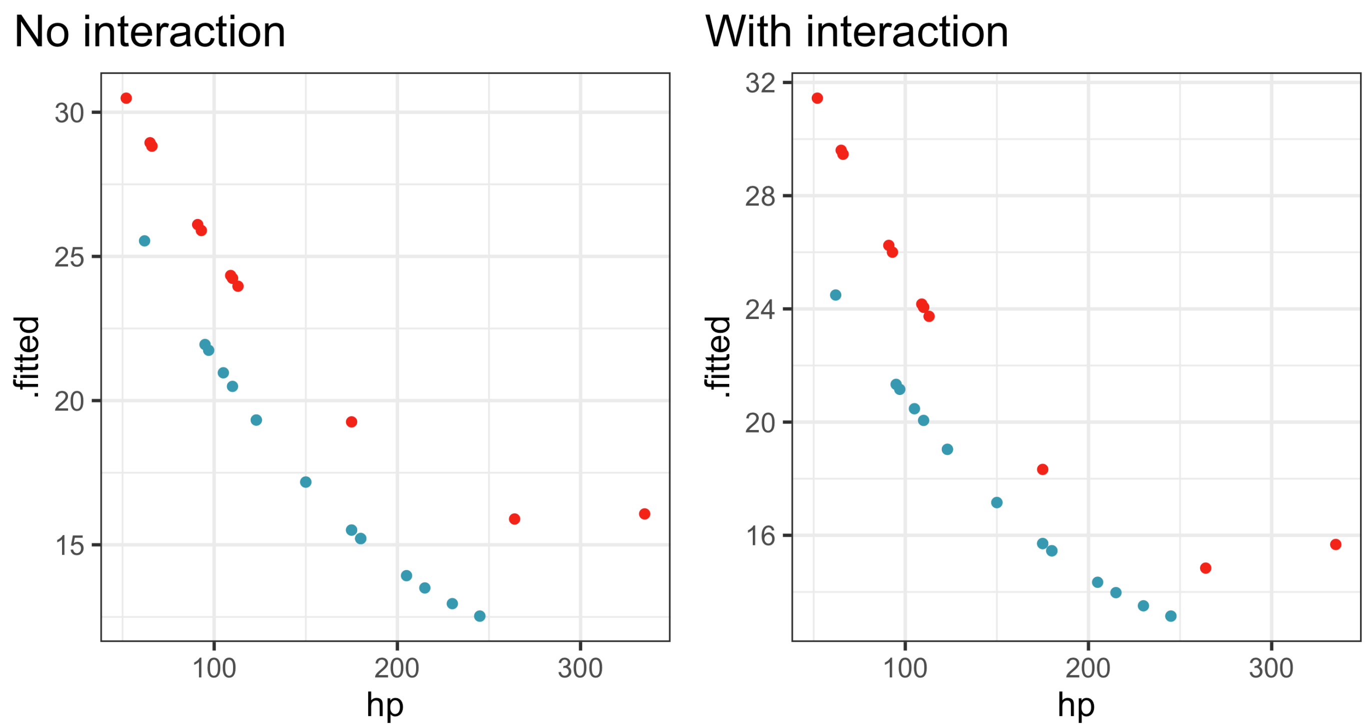
▶ Code



Multiple variables

Missing terms





▶ Code

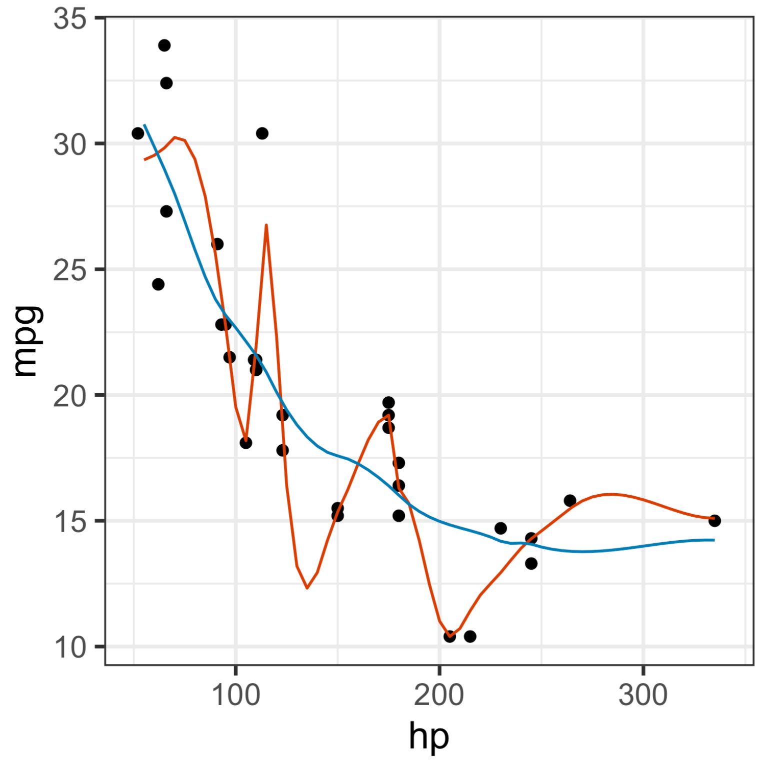
```
# A tibble: 4 × 5
  term          estimate std.error statistic  p.value
  <chr>            <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)      18.6      0.669     27.8  6.37e-22
2 poly(hp, 2)1    -23.5      2.79      -8.43 3.58e- 9
3 poly(hp, 2)2      7.88     3.14       2.51 1.80e- 2
4 ammanual          3.75     1.16       3.22 3.20e- 3

# A tibble: 6 × 5
  term                  estimate std.error statistic  p.value
  <chr>                    <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)              18.4      0.784    23.5   5.05e-19
2 poly(hp, 2)1            -20.4      5.02     -4.07  3.93e- 4
3 poly(hp, 2)2              7.02     7.09      0.990 3.32e- 1
4 ammanual                  3.55     1.22      2.92  7.11e- 3
5 poly(hp, 2)1:ammanu…     -5.97     6.36     -0.940 3.56e- 1
6 poly(hp, 2)2:ammanu…      2.93     8.12      0.361 7.21e- 1
```

Non-parametric model

Smoothing splines

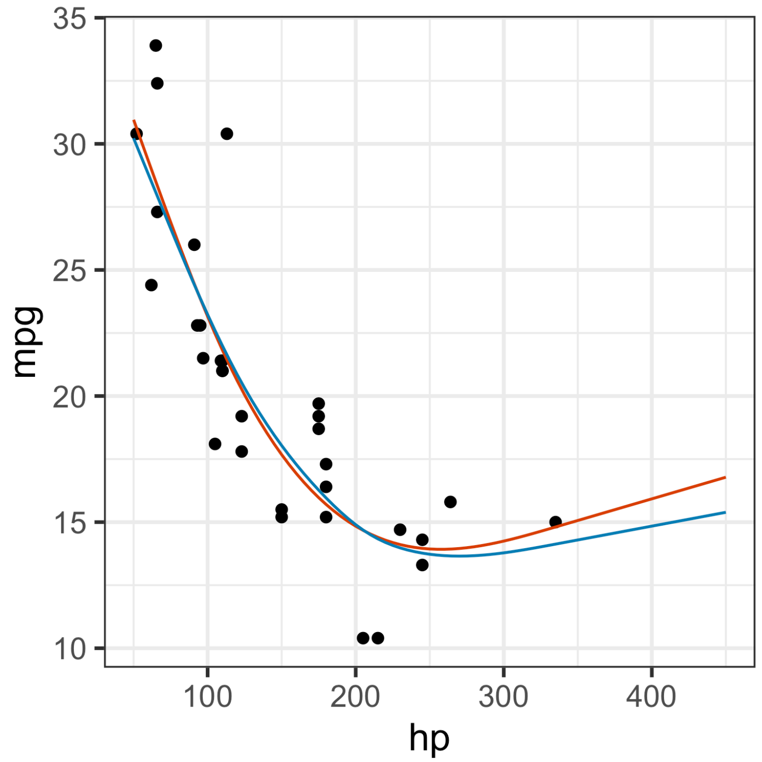
We’ve seen `loess`, which fits a linear model in a sliding window over predictor, where `span` controls size of window.

▶ Code



Smoothing splines, provide more advanced technique, and stability.

▶ Code



And are used to fit non-linear models to multiple predictors.

Logistic regression

- Not all parametric models assume normally distributed errors nor continuous responses.
- Logistic regression models the relationship between a set of explanatory variables $(x_{i1}, \ldots, x_{ik})$ and a set of **binary outcomes** Y_i for $i = 1, \ldots, n$.
- We assume that $Y_i \sim B(1, p_i) \equiv \text{Bernoulli}(p_i)$ and the model is given by

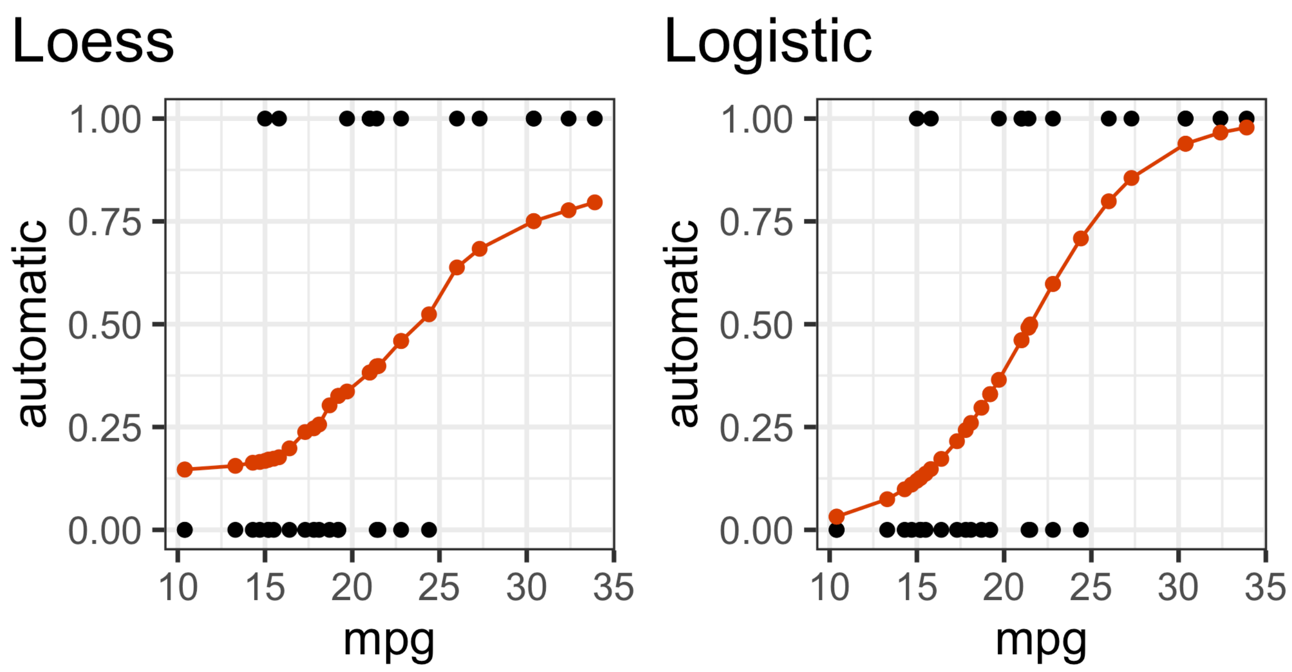
$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}.$$

- Taking the exponential of both sides and rearranging we get

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})}}.$$

- The function $f(p) = \ln\left(\frac{p}{1-p}\right)$ is called the **logit** function, continuous with range $(-\infty, \infty)$, and if p is the probablity of an event, $f(p)$ is the log of the odds.

▶ Code



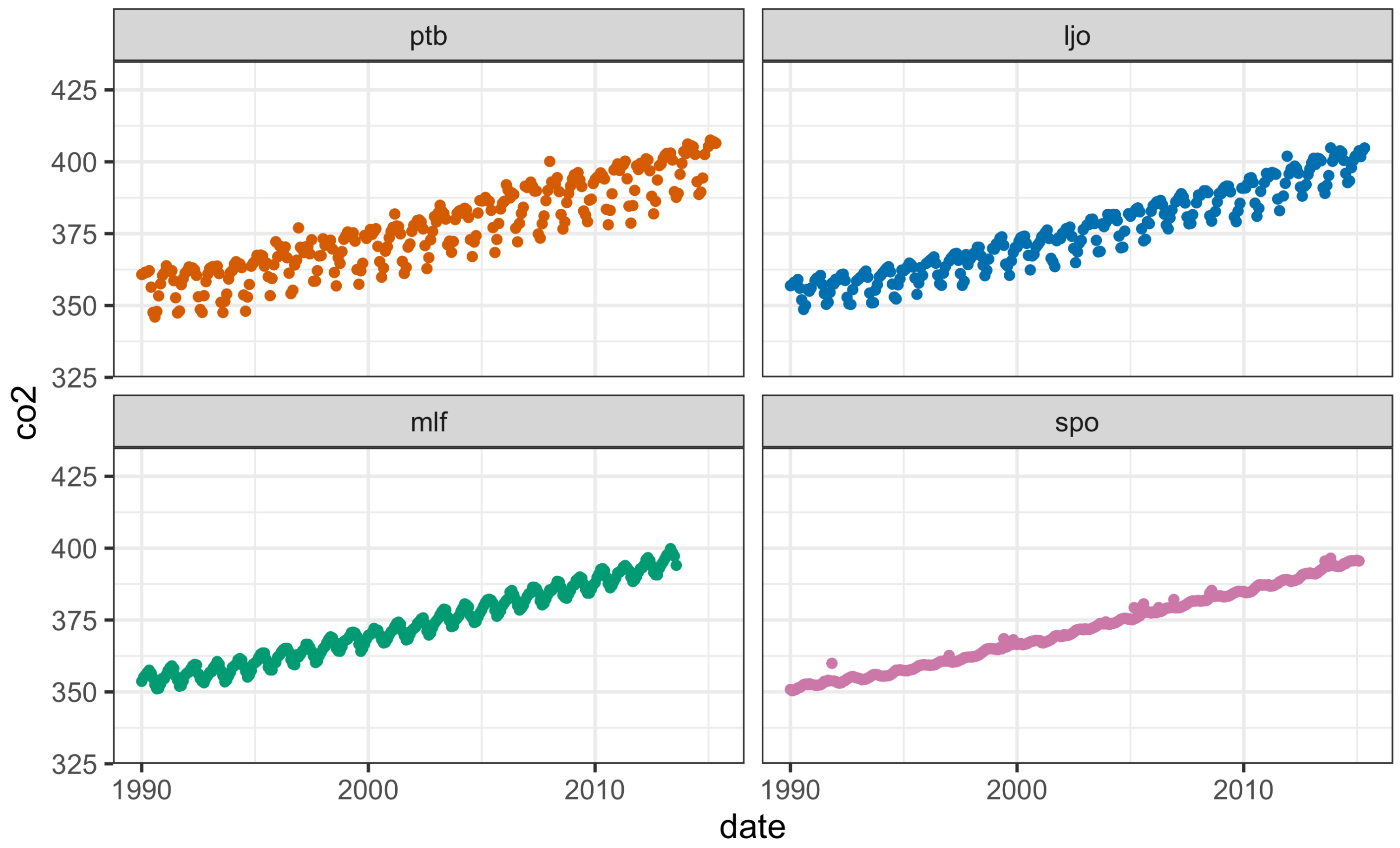
Slide a window and compute average (proportion) using loess, vs logistic function.

Time series

Trend and seasonality

▶ Code

Data | Trend | Seasonality



Data from Global CO2 monitoring stations

Exploring lags

Melbourne’s temperature, from high to low!

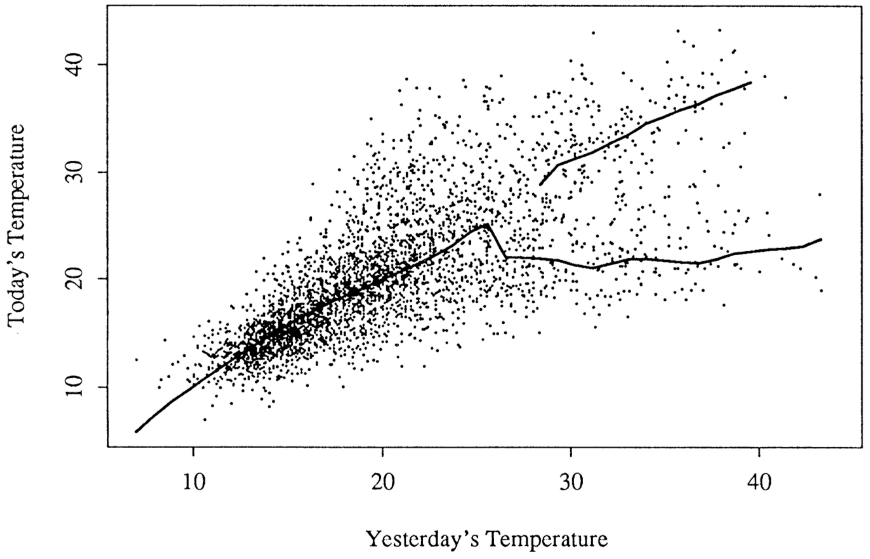


Figure 1. A Lagged Scatterplot of Each Day's Temperature Against the Previous Day's Temperature. Note the two "arms" on the right of the plot. The lines shown are from a modal regression discussed in Section 5.3.

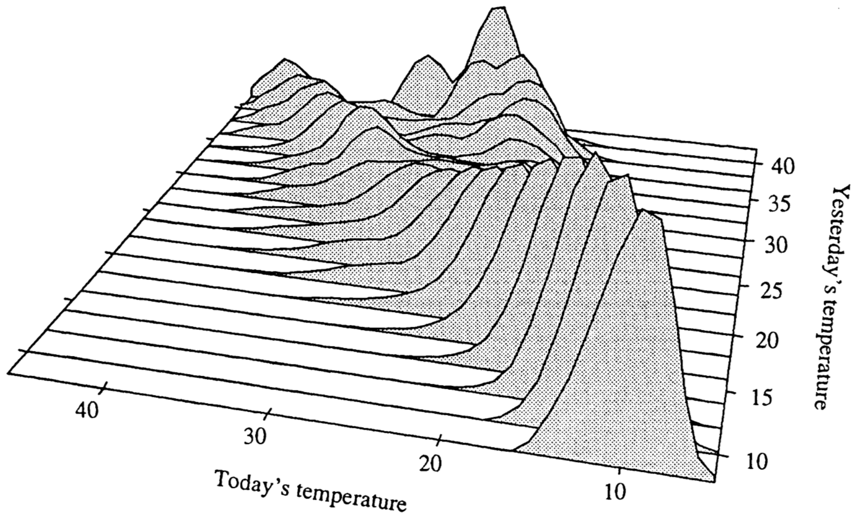


Figure 2. Stacked Conditional Density Plot of Temperature Conditional on the Previous Day's Temperature. The bimodality of the distribution of temperature following a hot day is more clear here than in Figure 1.

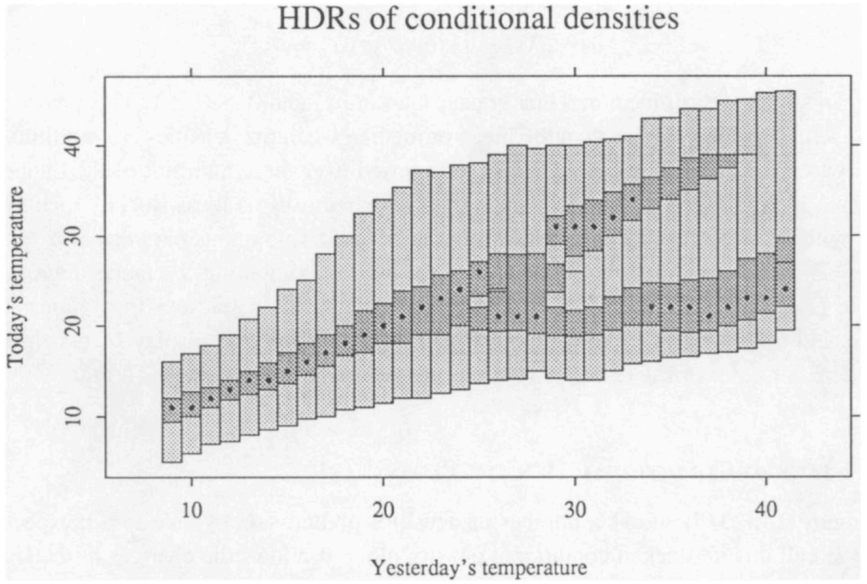


Figure 4. Highest Density Regions (50% and 99%) for Maximum Daily Temperature Conditional on the Previous Day's Maximum Temperature. Conditional modes are also marked (by •) for each x value. Compare this plot with the scatterplot of Figure 1 and the modal regression plot of Figure 5.

Today plotted vertically, yesterday plotted horizontally. Different types of plots are different models applied to the data (lags).

Hyndman, Bashtannyk, Grunwald (1996)

High-dimensions

Groups

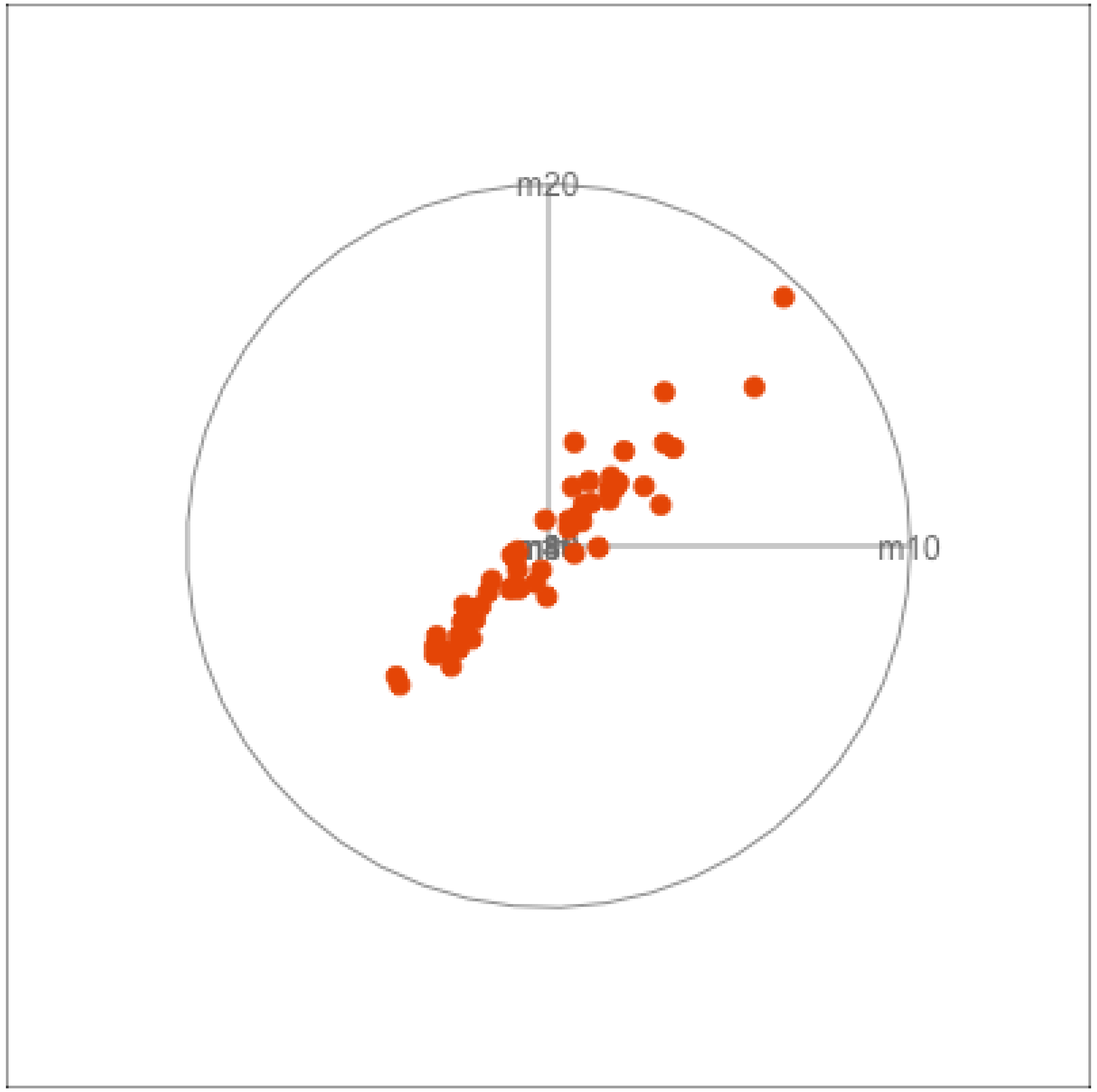
Data

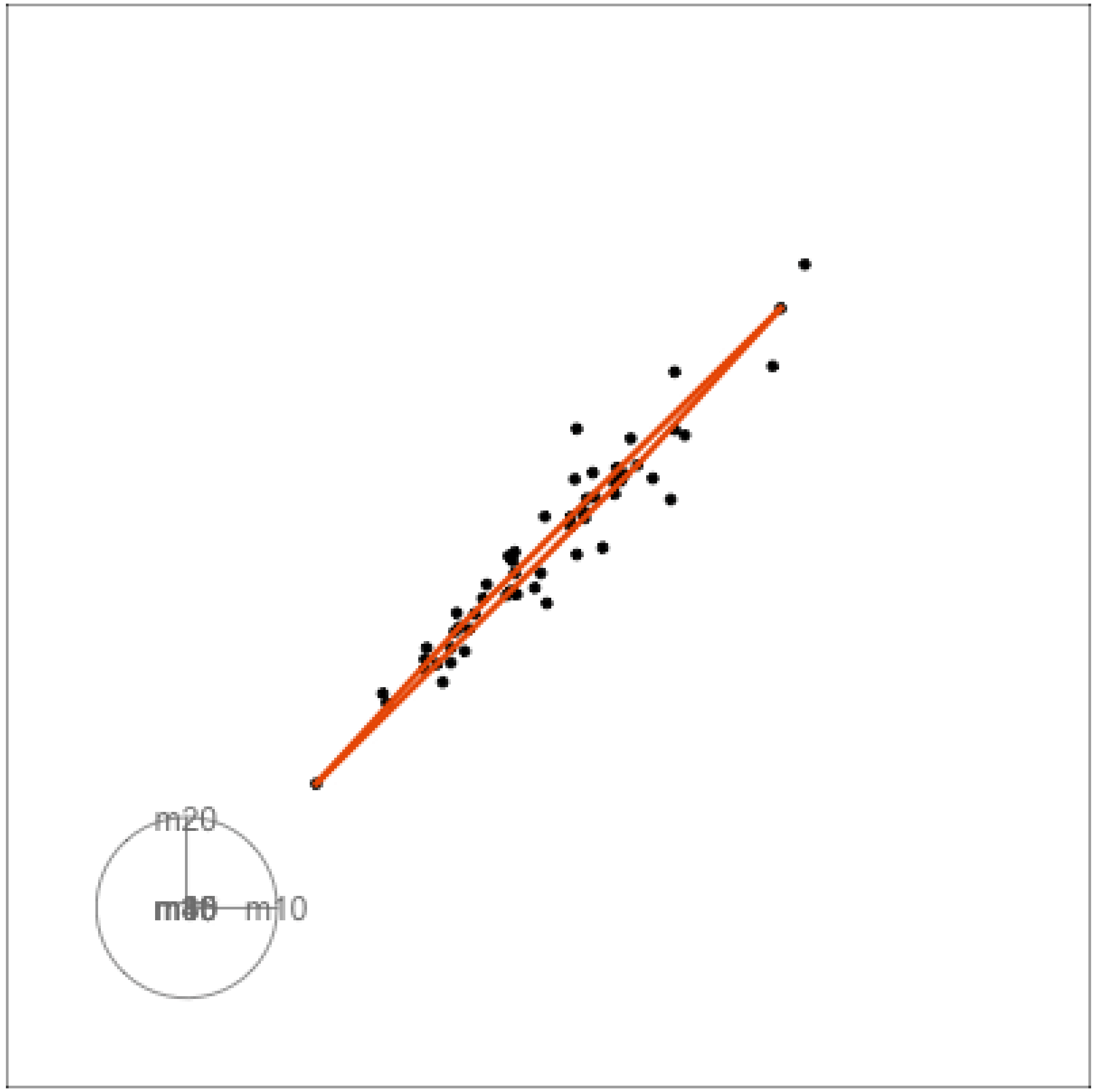
A little fuzzy.

Model view

Clearer view. Misses some quirks.

Relationships





Take-aways

- Models provide different lenses for extracting the patterns in the data
  - Sharpen
  - Exaggerate
  - Hallucinate
- Form a decomposition of the observed values into different strata
- Provide a multitude of other numerical quantities with which to see various aspects of the data.
- We are already using models, all the time, when making plots.

Resources

- Cook & Weisberg (1994) An Introduction to Regression Graphics
- Belsley, Kuh and Welsch (1980). Regression Diagnostics
- Hyndman, Bashtannyk, Grunwald (1996) Estimating and Visualizing Conditional Densities