Case study 1: the windmill data

The windmill data

- Engineer: does amount of electricity generated by windmill depend on how strongly wind blowing?
- Measurements of wind speed and DC current generated at various times.
- Assume the "various times" to be randomly selected — aim to generalize to "this windmill at all times".
- Research questions:
  - Relationship between wind speed and current generated?
  - If so, what kind of relationship?
  - Can we model relationship to do predictions?

Packages for this section

```
library(tidyverse)
library(broom)
```

Reading in the data

```
my_url <- "http://www.utsc.utoronto.ca/~butler/c32/windmill.cs
windmill <- read_csv(my_url)

## Parsed with column specification:
## cols(
##   wind_velocity = col_double(),
##   DC_output = col_double()
## )
```

The data

```
windmill
```

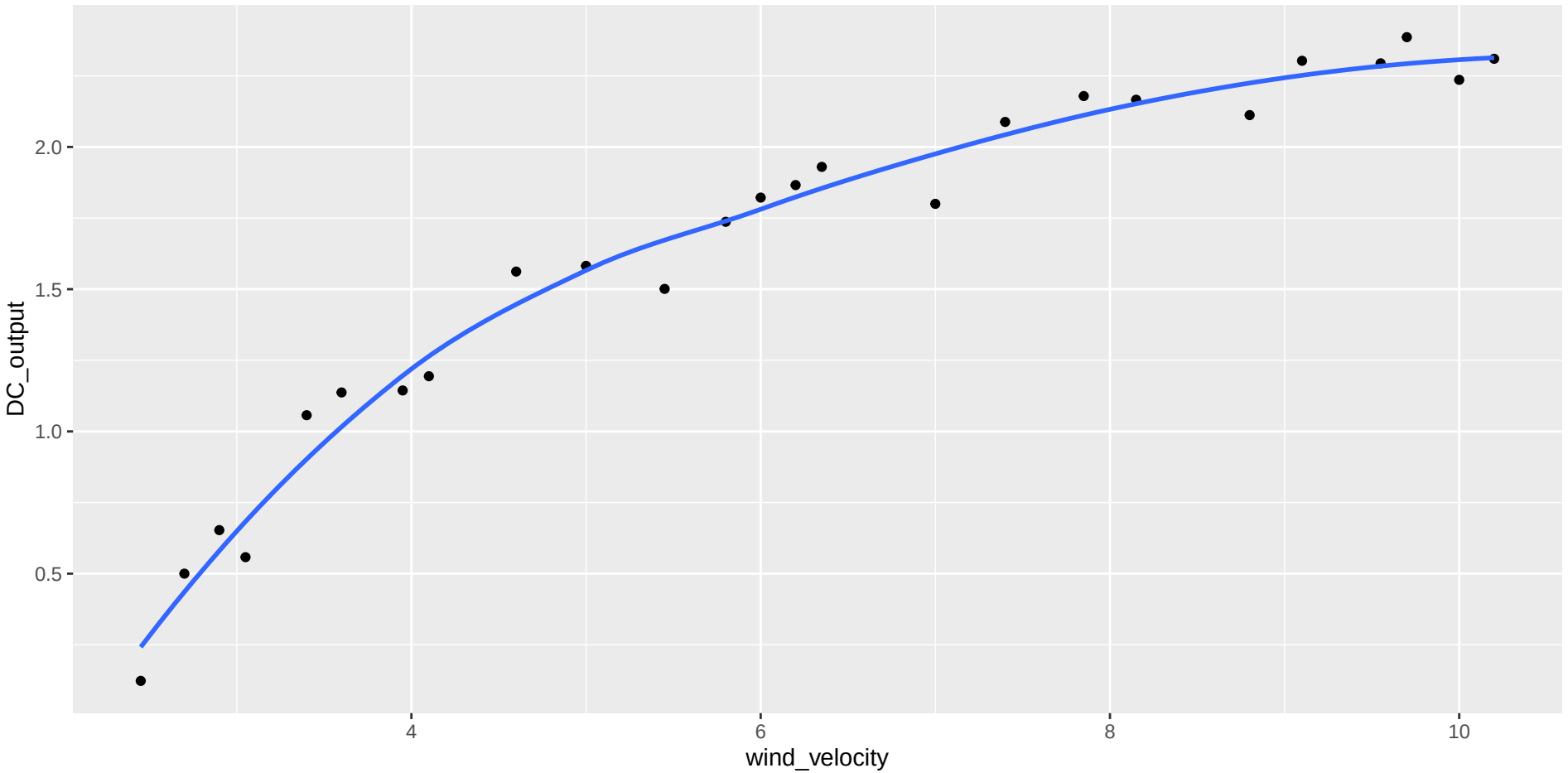
wind_velocity	DC_output
5.00	1.582
6.00	1.822
3.40	1.057
2.70	0.500
10.00	2.236
9.70	2.386
9.55	2.294
3.05	0.558
8.15	2.166
6.20	1.866
2.90	0.653
6.35	1.930
4.60	1.562

Strategy

- Two quantitative variables, looking for relationship: regression methods.
- Start with picture (scatterplot).
- Fit models and do model checking, fixing up things as necessary.
- Scatterplot:
  - 2 variables, `DC_output` and `wind_velocity`.
  - First is output/response, other is input/explanatory.
  - Put `DC_output` on vertical scale.
- Add trend, but don't want to assume linear:

```
ggplot(windmill, aes(y = DC_output, x = wind_velocity)) +
  geom_point() + geom_smooth(se = F)
```

Scatterplot



Comments

- Definitely a relationship: as wind velocity increases, so does DC output. (As you'd expect.)
- Is relationship linear? To help judge, `geom_smooth` smooths scatterplot trend. (Trend called "loess", "Locally weighted least squares" which downweights outliers. Not constrained to be straight.)
- Trend more or less linear for while, then curves downwards (levelling off?). Straight line not so good here.

Fit a straight line (and see what happens)

```
DC.1 <- lm(DC_output ~ wind_velocity, data = windmill)
summary(DC.1)
```

```
##
## Call:
## lm(formula = DC_output ~ wind_velocity, data = windmill)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59869 -0.14099  0.06059  0.17262  0.32184
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.13088    0.12599   1.039     0.31
## wind_velocity  0.24115    0.01905  12.659 7.55e-12 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2361 on 23 degrees of freedom
```

Another way of looking at the output

- The standard output tends to go off the bottom of the page rather easily. Package `broom` has these:

```
glance(DC.1)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC
0.8744932	0.8690364	0.2360521	160.2571	0	2	1.66138	2.67724

showing that the R-squared is 87%, and

```
tidy(DC.1)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.1308751	0.1259894	1.038779	0.3097053
wind_velocity	0.2411489	0.0190492	12.659268	0.0000000

showing the intercept and slope and their significance.

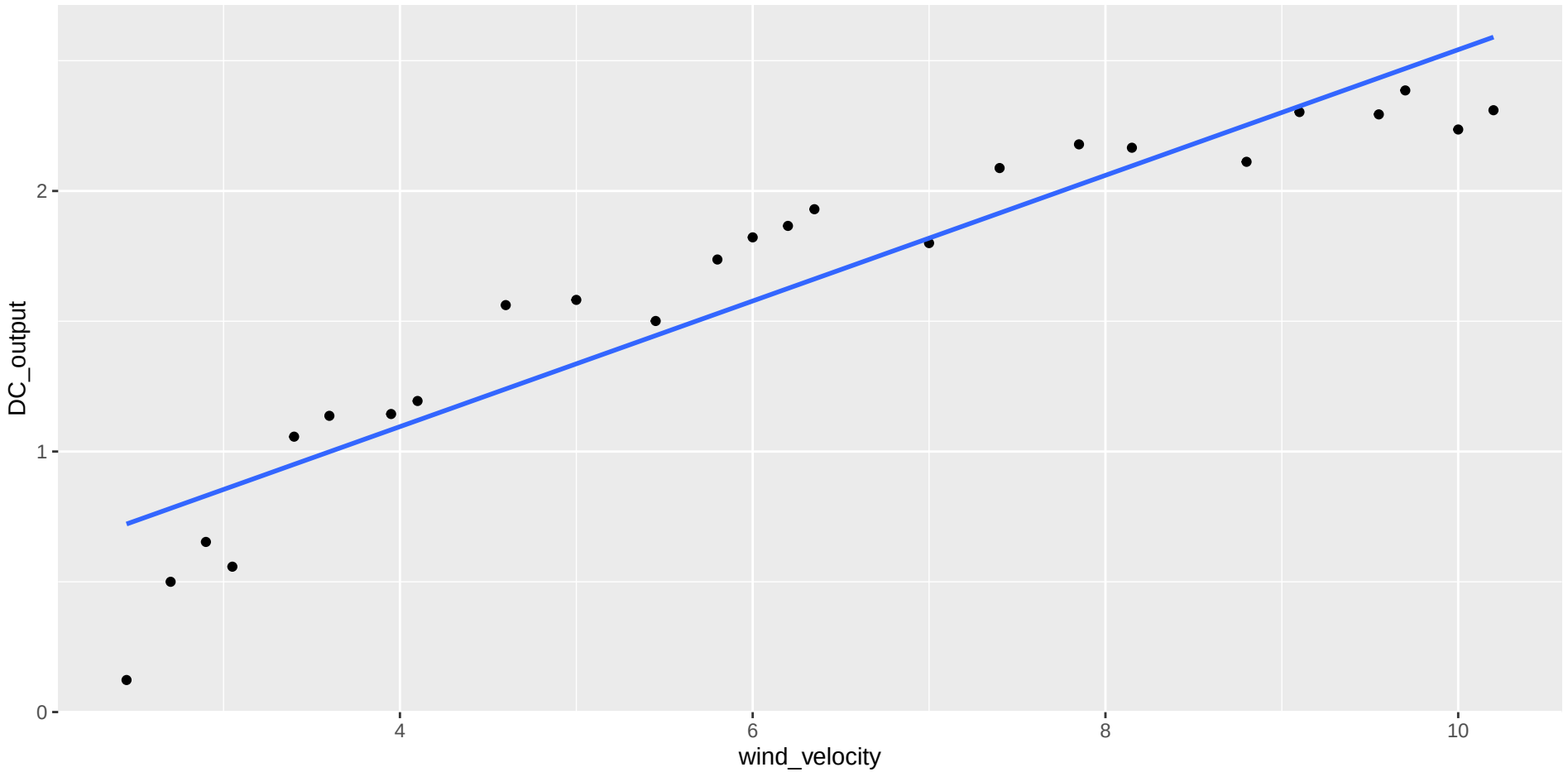
Comments

- Strategy: `lm` actually fits the regression. Store results in a variable. Then look at the results, eg. via `summary` or `glance`/`tidy`.
- My strategy for model names: base on response variable (or data frame name) and a number. Allows me to fit several models to same data and keep track of which is which.
- Results actually pretty good: `wind.velocity` strongly significant, R-squared (87%) high.
- How to check whether regression is appropriate? Look at the residuals, observed minus predicted, plotted against fitted (predicted).
- Plot using the regression object as "data frame" (see over).

back to scatterplot, but this time add line

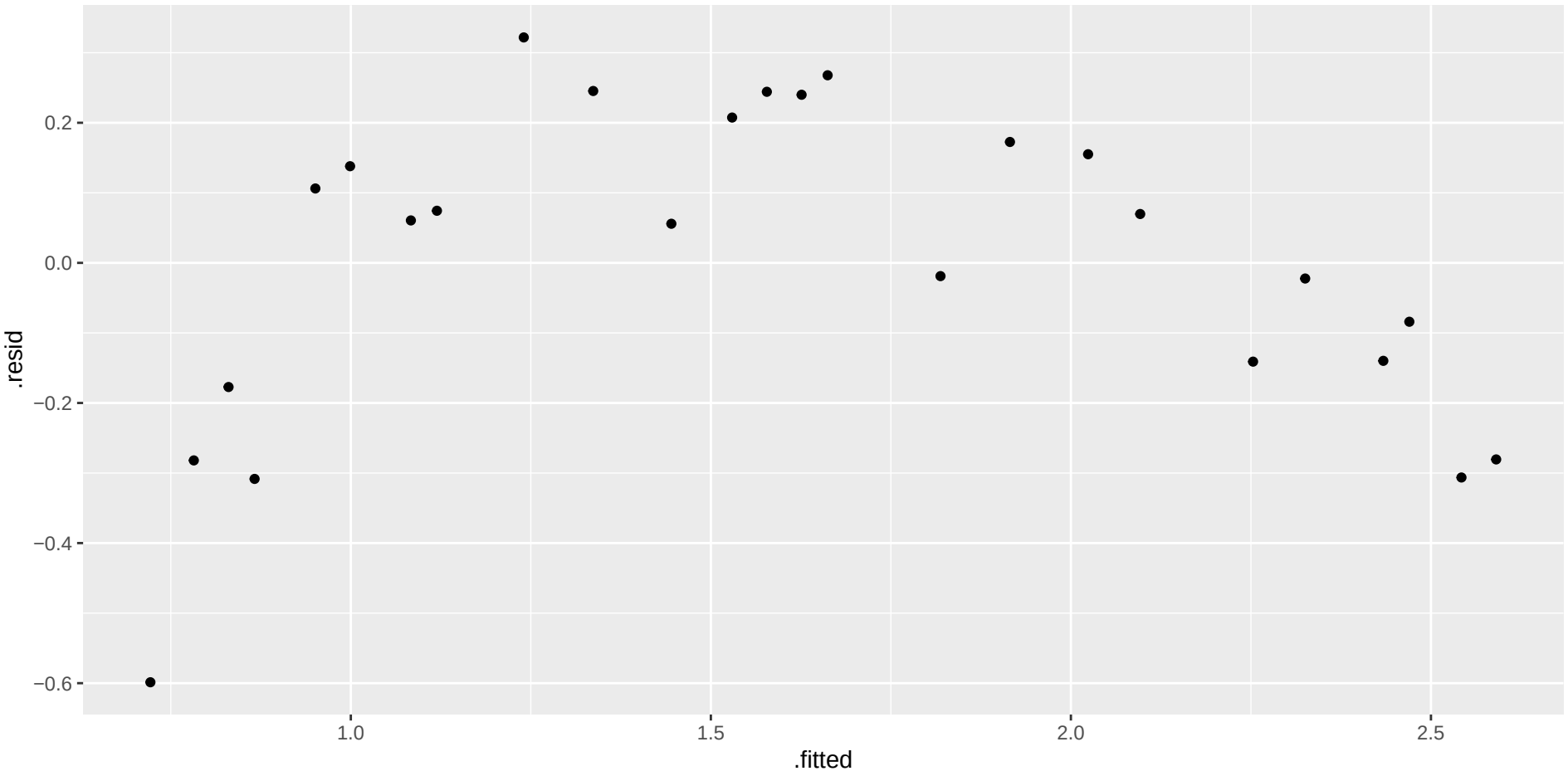
```
ggplot(windmill, aes(y = DC_output, x = wind_velocity)) +
  geom_point() + geom_smooth(method="lm", se = F)
```

The plot



Plot of residuals against fitted values

```
ggplot(DC.1, aes(y = .resid, x = .fitted)) + geom_point()
```



Comments on residual plot

- Residual plot should be a random scatter of points.
- Should be no pattern "left over" after fitting the regression.
- Smooth trend should be more or less straight across at 0.
- Here, have a curved trend on residual plot.
- This means original relationship must have been a curve (as we saw on original scatterplot).
- Possible ways to fit a curve:
  - Add a squared term in explanatory variable.
  - Transform response variable (doesn't work well here).
  - See what science tells you about mathematical form of relationship, and try to apply.

Parabolas and fitting parabola model

- A parabola has equation

$$y = ax^2 + bx + c$$

  with coefficients a, b, c. About the simplest function that is not a straight line.
- Fit one using `lm` by adding x^2 to right side of model formula with $+$:

```
DC.2 <- lm(DC_output ~ wind_velocity + I(wind_velocity^2),
  data = windmill
)
```

- The `I()` necessary because `^` in model formula otherwise means something different (to do with interactions in ANOVA).
- This actually *multiple regression*.
- Call it *parabola model*.

Parabola model output

```
tidy(DC.2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-1.1558982	0.1746500	-6.618368	1.2e-06
wind_velocity	0.7229359	0.0614254	11.769340	0.0e+00
I(wind_velocity^2)	-0.0381209	0.0047967	-7.947349	1.0e-07

```
glance(DC.2)
```

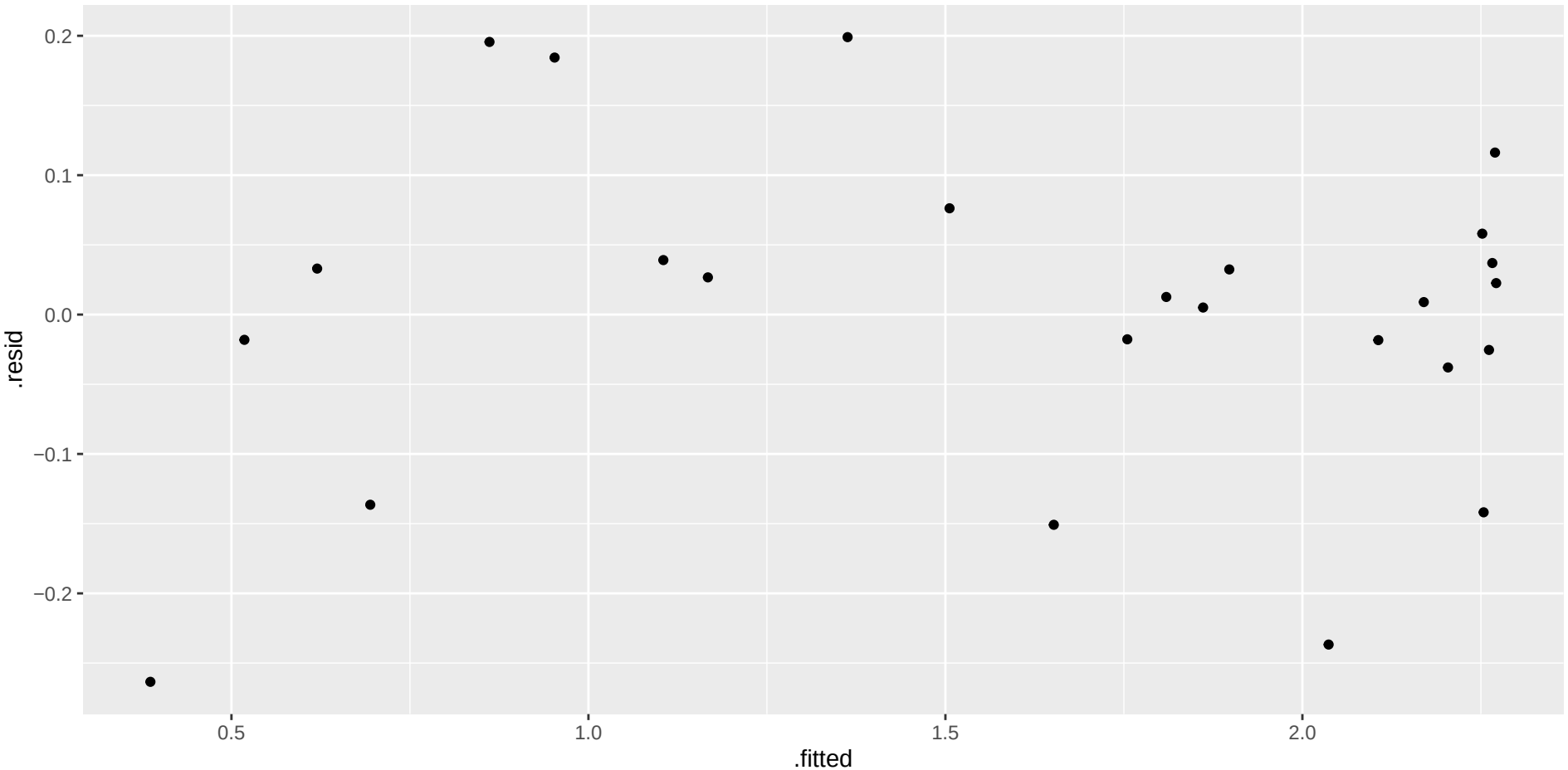
r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik
0.9675771	0.9646295	0.1226741	328.266	0	3	18.58005

Comments on output

- R-squared has gone up a lot, from 87% (line) to 97% (parabola).
- Coefficient of squared term strongly significant (P-value 6.59×10^{-8}).
- Adding squared term has definitely improved fit of model.
- Parabola model better than linear one.
- But...need to check residuals again.

Residual plot from parabola model

```
ggplot(DC.2, aes(y = .resid, x = .fitted)) +
  geom_point()
```



Make scatterplot with fitted line and curve

- Residual plot basically random. Good.
- Scatterplot with fitted line and curve like this:

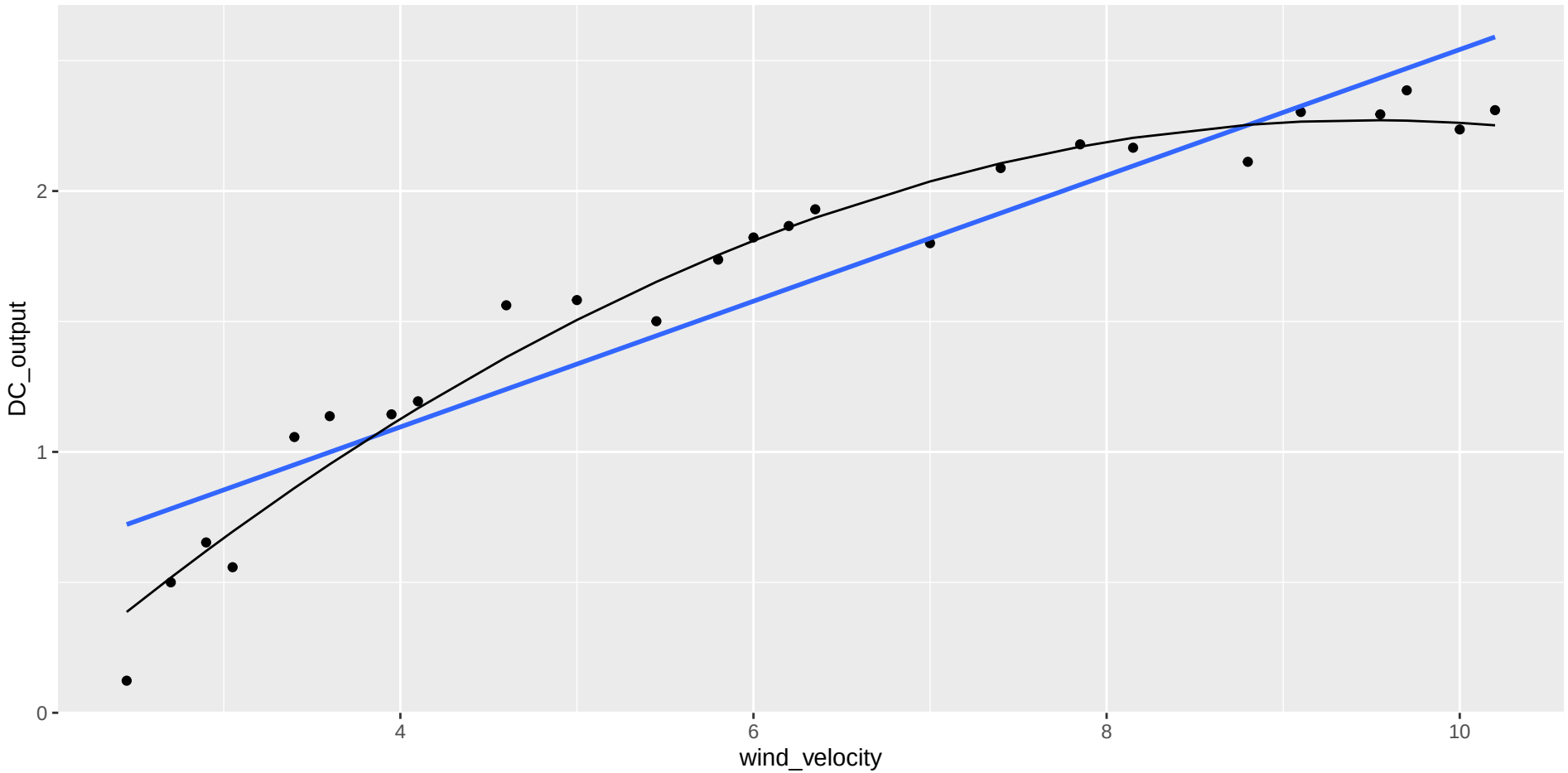
```
ggplot(windmill, aes(y = DC_output, x = wind_velocity)) +
  geom_point() + geom_smooth(method = "lm", se = F) +
  geom_line(data = DC.2, aes(y = .fitted))
```

Comments

- This plots:
  - scatterplot (`geom_point`);
  - straight line (via tweak to `geom_smooth`, which draws best-fitting line);
  - fitted curve, using the predicted `DC_output` values, joined by lines (with points not shown).
- Trick in the `geom_line` is use the predictions as the `y`-points to join by lines (from `DC.2`), instead of the original data points. Without the `data` and `aes` in the `geom_line`, original data points would be joined by lines.

Scatterplot with fitted line and curve

```
## `geom_smooth()` using formula 'y ~ x'
```



Curve clearly fits better than line.

Another approach to a curve

- There is a problem with parabolas, which we'll see later.
- Go back to engineer with findings so far. Ask, "what should happen as wind velocity increases?":
  - Upper limit on electricity generated, but otherwise, the larger the wind velocity, the more electricity generated.
- Mathematically, sounds like asymptote. Straight lines and parabolas don't have them, but eg. $y = 1/x$ does: as x gets bigger, y approaches zero without reaching it.
- What happens to $y = a + b(1/x)$ as x gets large?
  - y gets closer and closer to a: that is, a is asymptote.
- Fit this, call it asymptote model.
- Fitting the model here because we have math to justify it.
- Alternative $y = a + bc^{-x}$ approaches asymptote faster

How to fit asymptote model?

- Define new explanatory variable to be $1/x$, and predict y from it.
- x is velocity, distance over time.
- So $1/x$ is time over distance. In walking world, if you walk 5 km/h, take 12 minutes to walk 1 km, called your pace. So 1 over `wind_velocity` we call `wind_pace`.
- Make a scatterplot first to check for straightness (next page).

```
windmill %>% mutate(wind_pace = 1 / wind_velocity) -> windmill
ggplot(windmill, aes(y = DC_output, x = wind_pace)) +
  geom_point() + geom_smooth(se = F)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

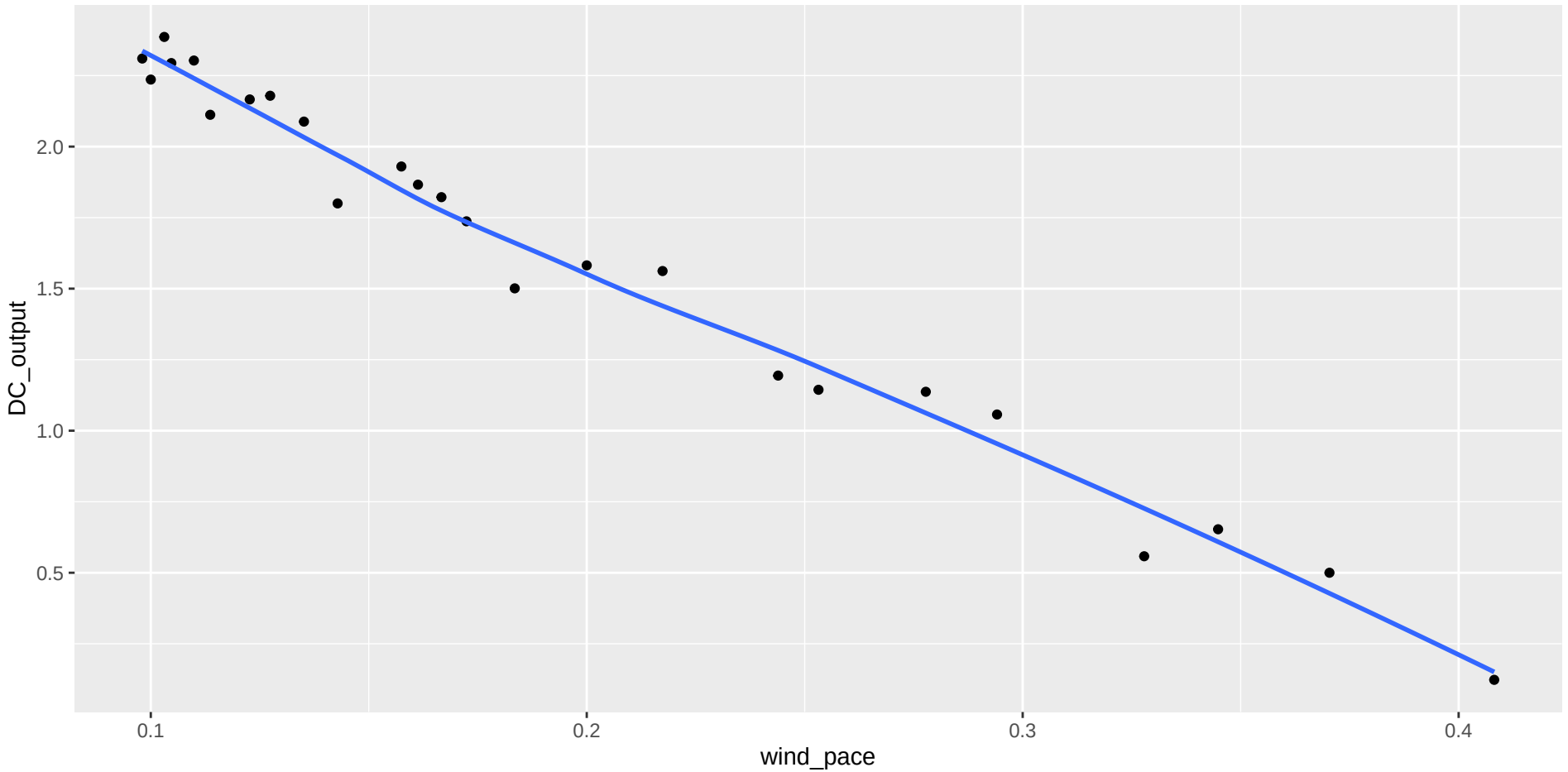
- and run regression like this (output page after):

```
DC.3 <- lm(DC_output ~ wind_pace, data = windmill)
```

Scatterplot for wind_pace

Pretty straight. Blue actually smooth curve not line:

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Regression output

```
glance(DC.3)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik
0.9800249	0.9791564	0.0941714	1128.433	0	2	24.63479

```
tidy(DC.3)
```

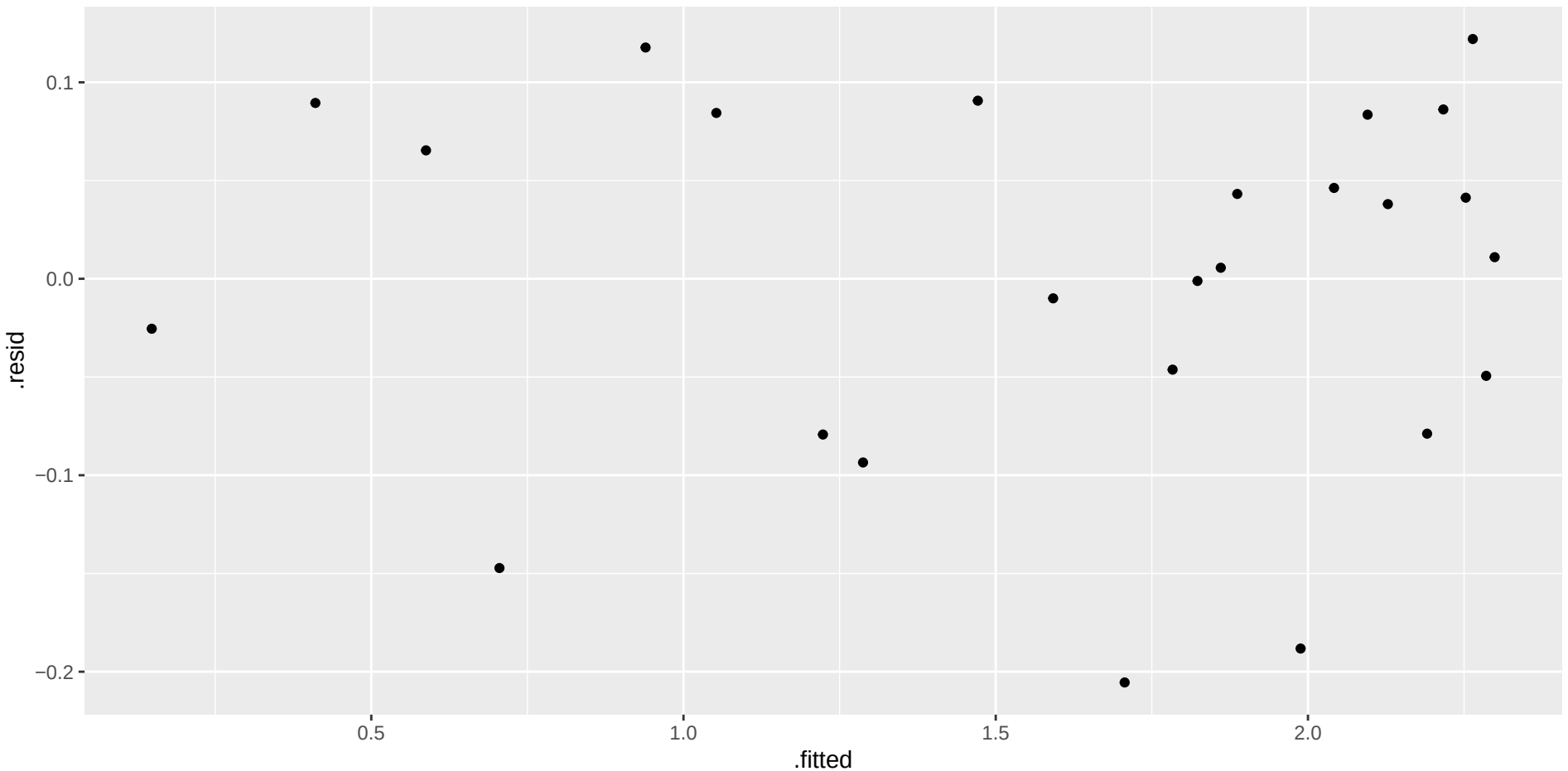
term	estimate	std.error	statistic	p.value
(Intercept)	2.978860	0.0449023	66.34096	0
wind_pace	-6.934547	0.2064336	-33.59215	0

Comments

- R-squared, 98%, even higher than for parabola model (97%).
- Simpler model, only one explanatory variable (`wind.pace`) vs. 2 for parabola model (`wind.velocity` and its square).
- `wind.pace` (unsurprisingly) strongly significant.
- Looks good, but check residual plot (over).

Residual plot for asymptote model

```
ggplot(DC.3, aes(y = .resid, x = .fitted)) + geom_point()
```



Plotting trends on scatterplot

- Residual plot not bad. But residuals go up to 0.10 and down to −0.20, suggesting possible skewness (not normal). I think it's not perfect, but OK overall.
- Next: plot scatterplot with all three fitted lines/curves on it (for comparison), with legend saying which is which.
- First make data frame containing what we need, taken from the right places:

```
w2 <- tibble(
  wind_velocity = windmill$wind_velocity,
  DC_output = windmill$DC_output,
  linear = fitted(DC.1),
  parabola = fitted(DC.2),
  asymptote = fitted(DC.3)
)
```

What's in `w2`

```
w2
```

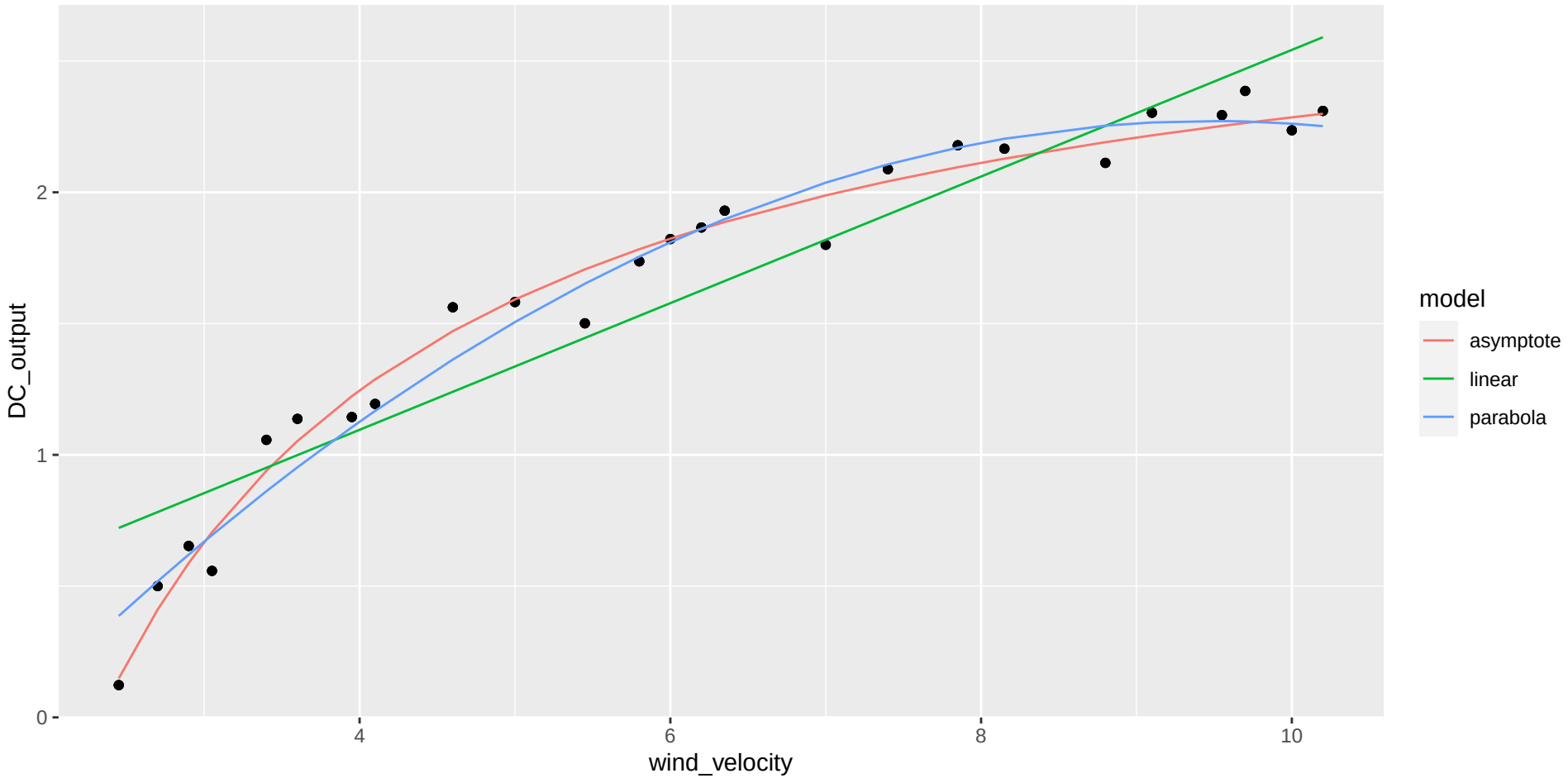
wind_velocity	DC_output	linear	parabola	asymptote
5.00	1.582	1.3366195	1.5057592	1.5919507
6.00	1.822	1.5777683	1.8093653	1.8231023
3.40	1.057	0.9507813	0.8614064	0.9392874
2.70	0.500	0.7819771	0.5181275	0.4105093
10.00	2.236	2.5423638	2.2613723	2.2854054
9.70	2.386	2.4700192	2.2697860	2.2639584
9.55	2.294	2.4338468	2.2714197	2.2527296
3.05	0.558	0.8663792	0.6944367	0.7052381
8.15	2.166	2.0962384	2.2039449	2.1279955
6.20	1.866	1.6259981	1.8609376	1.8603848
2.90	0.653	0.8302069	0.6200192	0.5876369
6.35	1.930	1.6621705	1.8976154	1.8868055
4.60	1.562	1.2401599	1.3629690	1.4713499

Making the plot

- ggplot likes to have one column of x's to plot, and one column of y's, with another column for distinguishing things.
- But we have three columns of fitted values, that need to be combined into one.
- pivot_longer, then plot:

```
w2 %>%
  pivot_longer(linear:asymptote, names_to="model", values_to="
  ggplot(aes(x = wind_velocity, y = DC_output)) +
  geom_point() +
  geom_line(aes(y = fit, colour = model))
```

Scatterplot with fitted curves



Comments

- Predictions from curves are very similar.
- Predictions from asymptote model as good, and from simpler model (one x not two), so prefer those.
- Go back to asymptote model summary.

Asymptote model summary

```
tidy(DC.3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	2.978860	0.0449023	66.34096	0
wind_pace	-6.934547	0.2064336	-33.59215	0

Comments

- Intercept in this model about 3.
- Intercept of asymptote model is the asymptote (upper limit of `DC.output`).
- Not close to asymptote yet.
- Therefore, from this model, wind could get stronger and would generate appreciably more electricity.
- This is extrapolation! Would like more data from times when `wind.velocity` higher.
- Slope -7. Why negative?
  - As wind.velocity increases,
  - wind.pace goes down,
  - and DC.output goes up. Check.
- Actual slope number hard to interpret.

Checking back in with research questions

- Is there a relationship between wind speed and current generated?
  - Yes.
- If so, what kind of relationship is it?
  - One with an asymptote.
- Can we model the relationship, in such a way that we can do predictions?
  - Yes, see model DC.3 and plot of fitted curve.
- Good. Job done.

Job done, kinda

- Just because the parabola model and asymptote model agree over the range of the data, doesn't necessarily mean they agree everywhere.
- Extend range of wind.velocity to 1 to 16 (steps of 0.5), and predict DC.output according to the two models:

```
wv <- seq(1, 16, 0.5)
wv
```

```
##  [1]  1.0  1.5  2.0  2.5  3.0  3.5  4.0  4.5  5.0
## [10]  5.5  6.0  6.5  7.0  7.5  8.0  8.5  9.0  9.5
## [19] 10.0 10.5 11.0 11.5 12.0 12.5 13.0 13.5 14.0
## [28] 14.5 15.0 15.5 16.0
```

- R has `predict`, which requires what to predict for, as data frame. The data frame has to contain values, with matching names, for all explanatory variables in regression(s).

Setting up data frame to predict from

- Linear model had just `wind_velocity`.
- Parabola model had that as well (squared one will be calculated)
- Asymptote model had just `wind_pace` (reciprocal of velocity).
- So create data frame called `wv_new` with those in:

```
wv_new <- tibble(wind_velocity = wv, wind_pace = 1 / wv)
```

wv_new

```
wv_new
```

wind_velocity	wind_pace
1.0	1.0000000
1.5	0.6666667
2.0	0.5000000
2.5	0.4000000
3.0	0.3333333
3.5	0.2857143
4.0	0.2500000
4.5	0.2222222
5.0	0.2000000
5.5	0.1818182
6.0	0.1666667
6.5	0.1538462
7.0	0.1428571

Doing predictions, one for each model

- Use same names as before:

```
linear <- predict(DC.1, wv_new)
parabola <- predict(DC.2, wv_new)
asymptote <- predict(DC.3, wv_new)
```

- Put it all into a data frame for plotting, along with original data:

```
my_fits <- tibble(
  wind_velocity = wv_new$wind_velocity,
  linear, parabola, asymptote
)
```

my_fits

```
my_fits
```

wind_velocity	linear	parabola	asymptote
1.0	0.3720240	-0.4710832	-3.9556872
1.5	0.4925984	-0.1572664	-1.6441714
2.0	0.6131729	0.1374900	-0.4884135
2.5	0.7337473	0.4131860	0.2050412
3.0	0.8543217	0.6698215	0.6673444
3.5	0.9748962	0.9073966	0.9975609
4.0	1.0954706	1.1259112	1.2452233
4.5	1.2160450	1.3253654	1.4378497
5.0	1.3366195	1.5057592	1.5919507
5.5	1.4571939	1.6670925	1.7180334
6.0	1.5777683	1.8093653	1.8231023
6.5	1.6983428	1.9325778	1.9120067
7.0	1.8189172	2.0367297	1.9882106

Making a plot 1/2

- To make a plot, we use the same trick as last time to get all three predictions on a plot with a legend (saving result to add to later):

```
my_fits %>%
    pivot_longer(
    linear:asymptote,
    names_to="model",
    values_to="fit"
  ) %>%
  ggplot(aes(
    y = fit, x = wind_velocity,
    colour = model
  )) + geom_line() -> g
```

Making a plot 2/2

- The observed wind velocities were in this range:

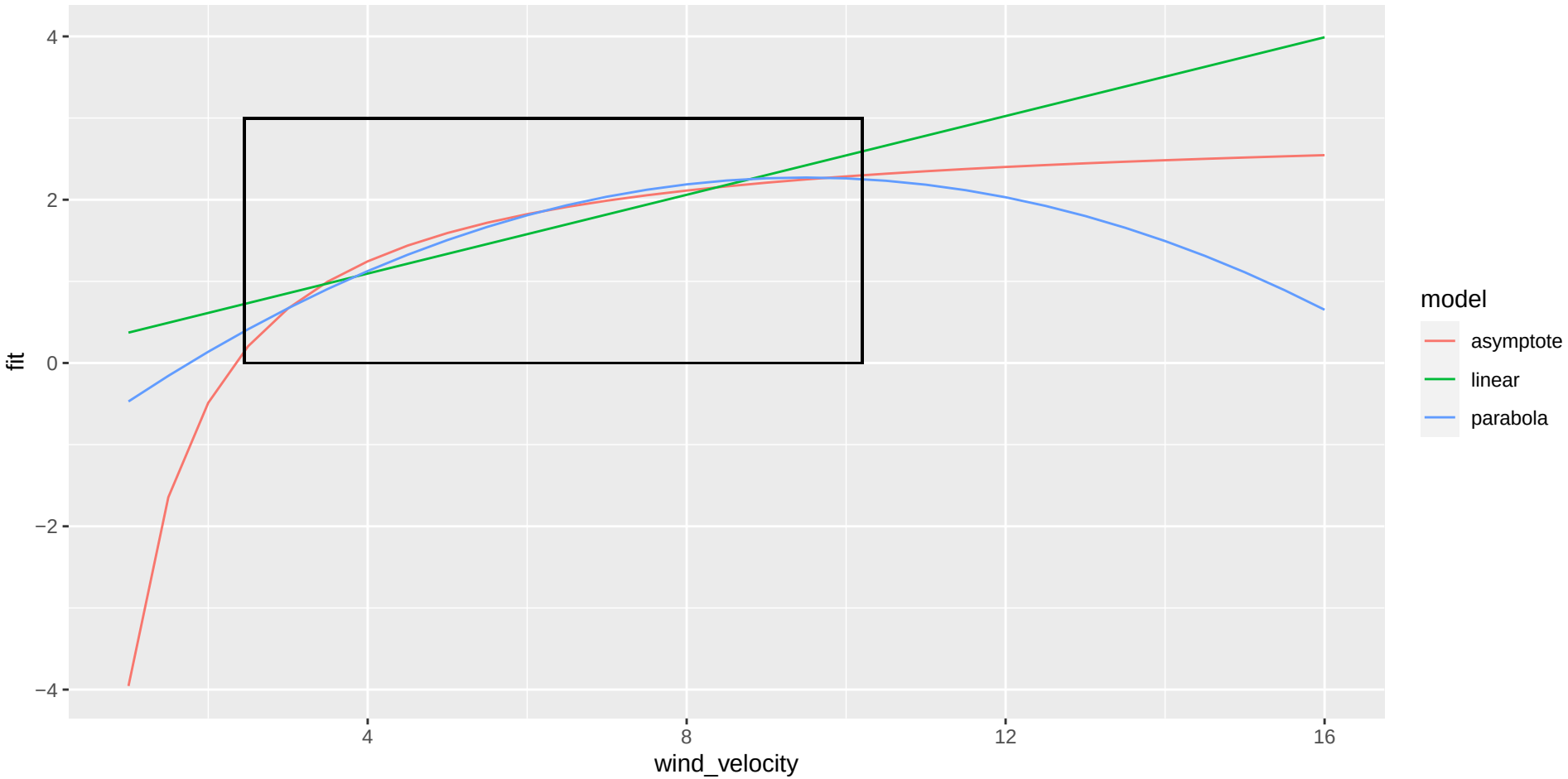
```
(vels <- range(windmill$wind_velocity))
```

```
## [1]  2.45 10.20
```

- `DC.output` between 0 and 3 from asymptote model. Add rectangle to graph around where the data were:

```
g + geom_rect(
  xmin = vels[1], xmax = vels[2], ymin = 0, ymax = 3,
  alpha = 0, colour = "black"
)
```

The plot


4
2
0
−2
−4
fit
4
8
12
16
wind_velocity
model
asymptote
linear
parabola

Comments (1)

- Over range of data, two models agree with each other well.
- Outside range of data, they disagree violently!
- For larger `wind.velocity`, asymptote model behaves reasonably, parabola model does not.
- What happens as `wind.velocity` goes to zero? Should find `DC.output` goes to zero as well. Does it?

Comments (2)

- For parabola model:

```
tidy(DC.2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-1.1558982	0.1746500	-6.618368	1.2e-06
wind_velocity	0.7229359	0.0614254	11.769340	0.0e+00
I(wind_velocity^2)	-0.0381209	0.0047967	-7.947349	1.0e-07

- Nope, goes to -1.16 (intercept), actually significantly different from zero.

Comments (3): asymptote model

```
tidy(DC.3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	2.978860	0.0449023	66.34096	0
wind_pace	-6.934547	0.2064336	-33.59215	0

- As `wind.velocity` heads to 0, wind.pace heads to $+\infty$, so DC.output heads to $-\infty$!
- Also need more data for small `wind.velocity` to understand relationship. (Is there a lower asymptote?)
- Best we can do now is to predict `DC.output` to be zero for small `wind.velocity`.
- Assumes a "threshold" wind velocity below which no electricity generated at all.

Summary

- Often, in data analysis, there is no completely satisfactory conclusion, as here.
- Have to settle for model that works OK, with restrictions.
- Always something else you can try.
- At some point you have to say "I stop."