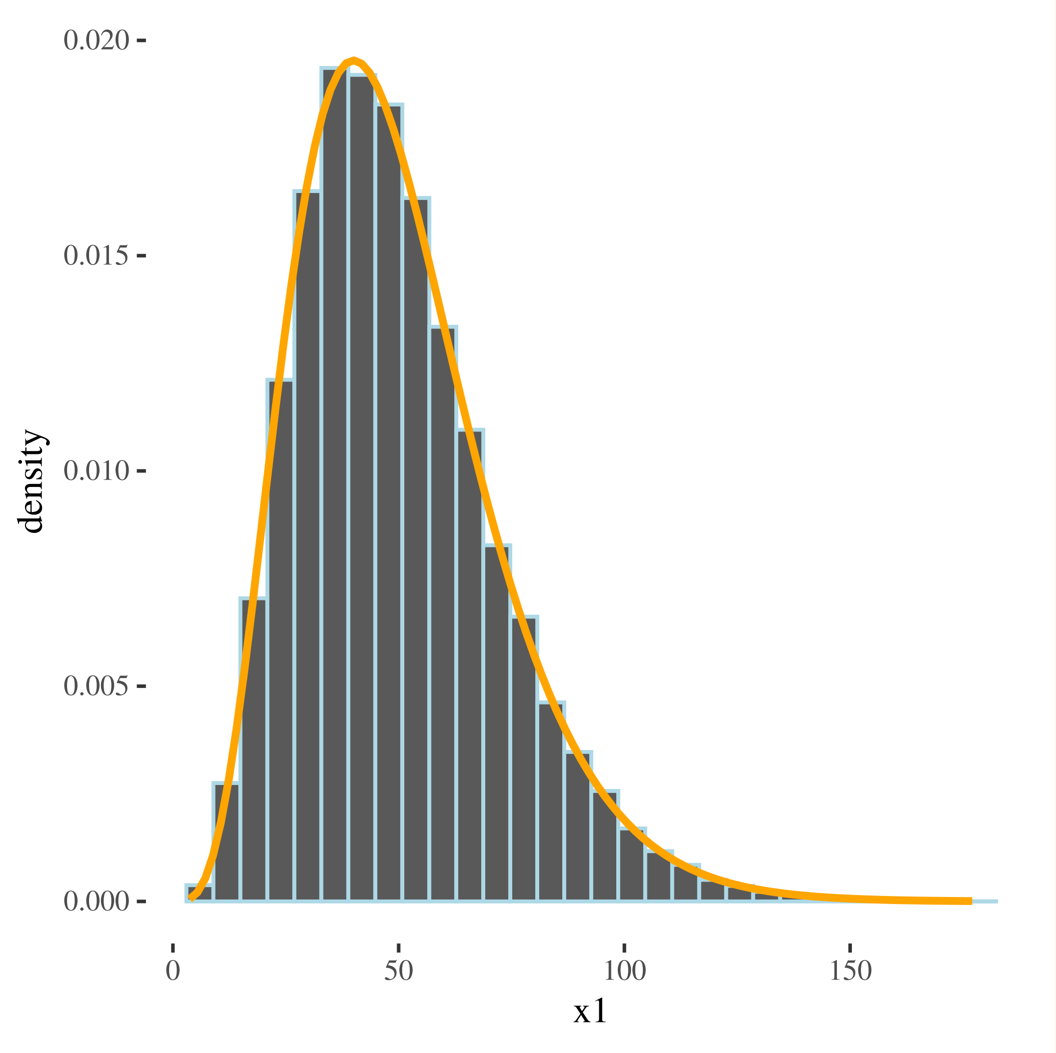
Poisson regression diagnostics

Stat 230

May 25 2022

Overview



Today:

Residuals and case influence

GOF and Poisson assumptions

Quasi-poisson model

Residuals

- Similar to **logistic regression**!
- Pearson and deviance: similar in values and pattern
- Plot vs predictors and look for a "null" plot
- When $\hat{\mu}_i$'s are large (at least 5), both types of residuals should be $N(0,1)$ distribution (approximately).

Pearson Residuals

Pearson residuals are basically response residuals standardized based on the Poisson SD:

$$pr_i = \frac{y_i - \hat{\mu}\left(X_i\right)}{\sqrt{\hat{\mu}\left(X_i\right)}}$$

- `resid(my_glm, type = "pearson")`
- `augment(my_glm, type.residuals = "pearson")`

Deviance Residuals

Deviance residuals are **each case's** contribution to the residual deviance:

$$\mathrm{Dres}_i = \mathrm{sign}(y_i - \hat{\mu}\left(X_i\right))\sqrt{2\left[y_i \ln\left(\frac{y_i}{\hat{\mu}\left(X_i\right)}\right) - (y_i - \hat{\mu}\left(X_i\right))\right]}$$

- `resid(my_glm, type(=)"deviance")`
- `augment(my_glm, type.residuals = "deviance")`

Case influence stats

In a GLM, **leverage** measures

- both a cases's "extremeness" in terms of it's predictor values and the size of a case's weight
- in a Poisson GLM, a case's weight is given by $\hat{\mu}\left(X_i\right)$

Cook's distance also takes into account a cases leverage (measured both by predictor values and it's estimated mean) and a case's residual value.

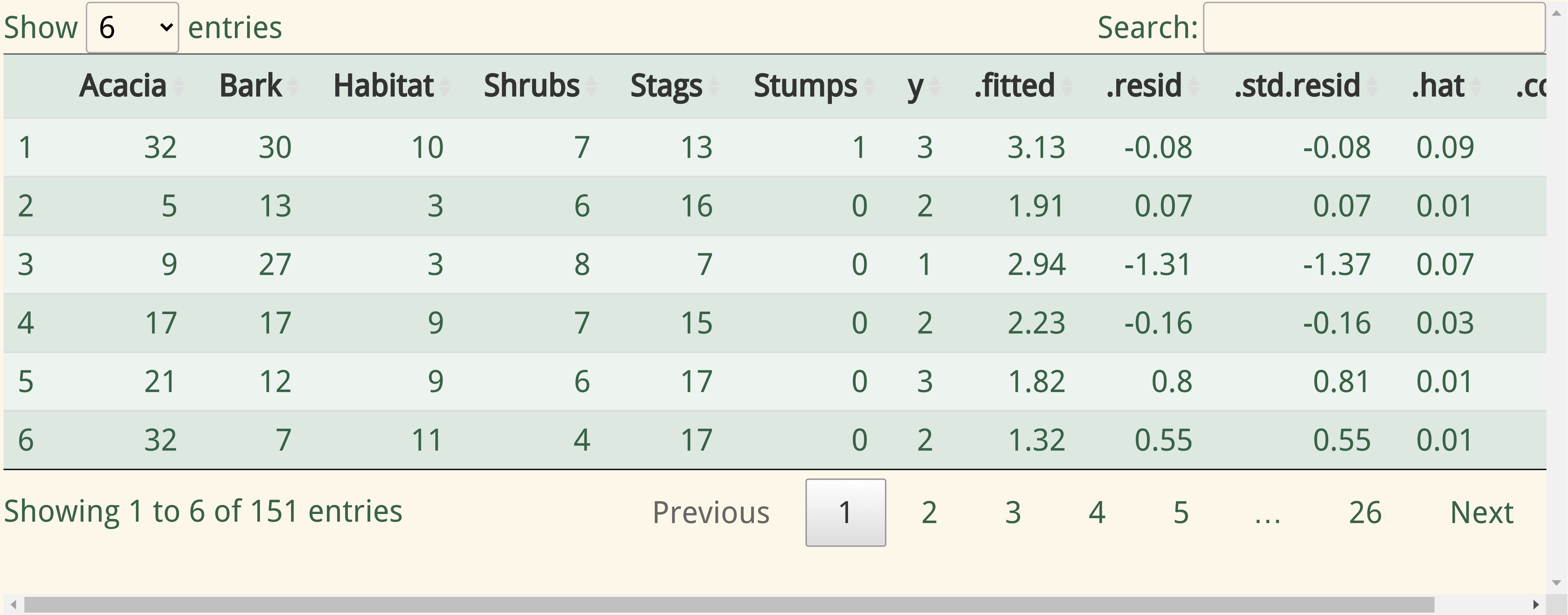
Australian Possums

```
possums <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/possums.csv
pos_glm <- glm(y ~ log(Bark), family = poisson, data = possums)
possums_aug <- augment(pos_glm, data=possums, type.predict = "response")
possums_aug_log <- augment(pos_glm, data=possums) # in log scale
summary(possums_aug$.fitted)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.4147  1.2033  1.4277  1.4768  1.6302  3.1325 
```

- Fitted values $\hat{\mu}$ are all less than 5 .
- residuals won't be approximately normal
- issues trusting GOF test

Augmented data



Show 6 entries

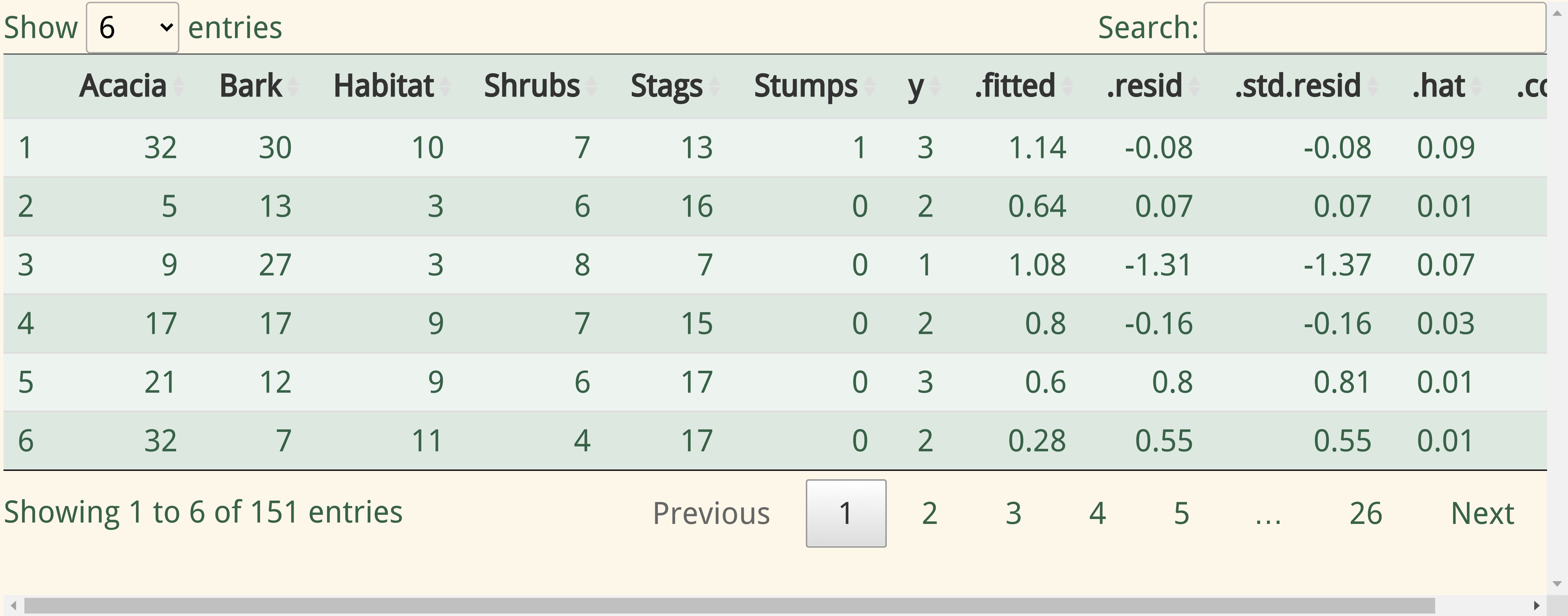
Search:

	Acacia	Bark	Habitat	Shrubs	Stags	Stumps	y	.fitted	.resid	.std.resid	.hat	.co
1	32	30	10	7	13	1	3	3.13	-0.08	-0.08	0.09	
2	5	13	3	6	16	0	2	1.91	0.07	0.07	0.01	
3	9	27	3	8	7	0	1	2.94	-1.31	-1.37	0.07	
4	17	17	9	7	15	0	2	2.23	-0.16	-0.16	0.03	
5	21	12	9	6	17	0	3	1.82	0.8	0.81	0.01	
6	32	7	11	4	17	0	2	1.32	0.55	0.55	0.01	

Showing 1 to 6 of 151 entries

Previous 1 2 3 4 5 … 26 Next

Augmented data (logged)



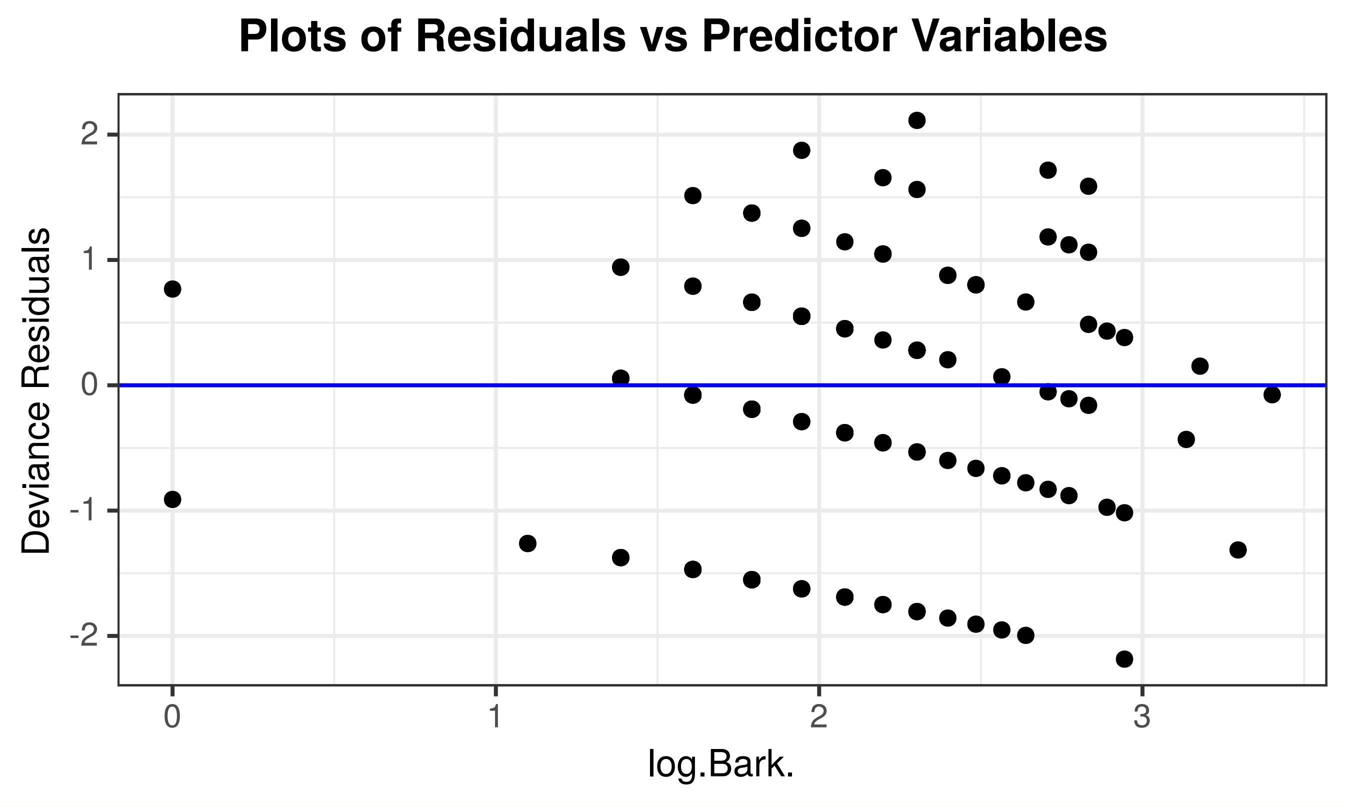
Show 6 entries

Search:

	Acacia	Bark	Habitat	Shrubs	Stags	Stumps	y	.fitted	.resid	.std.resid	.hat	.co
1	32	30	10	7	13	1	3	1.14	-0.08	-0.08	0.09	
2	5	13	3	6	16	0	2	0.64	0.07	0.07	0.01	
3	9	27	3	8	7	0	1	1.08	-1.31	-1.37	0.07	
4	17	17	9	7	15	0	2	0.8	-0.16	-0.16	0.03	
5	21	12	9	6	17	0	3	0.6	0.8	0.81	0.01	
6	32	7	11	4	17	0	2	0.28	0.55	0.55	0.01	

Showing 1 to 6 of 151 entries

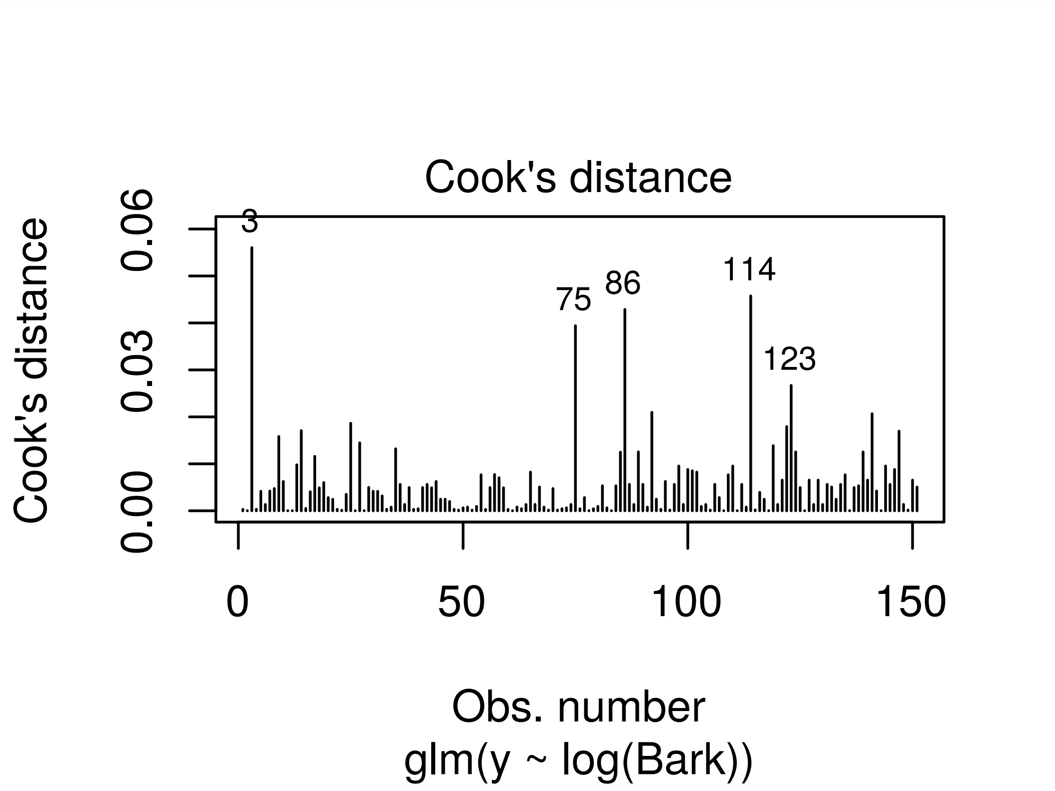
Australian Possums

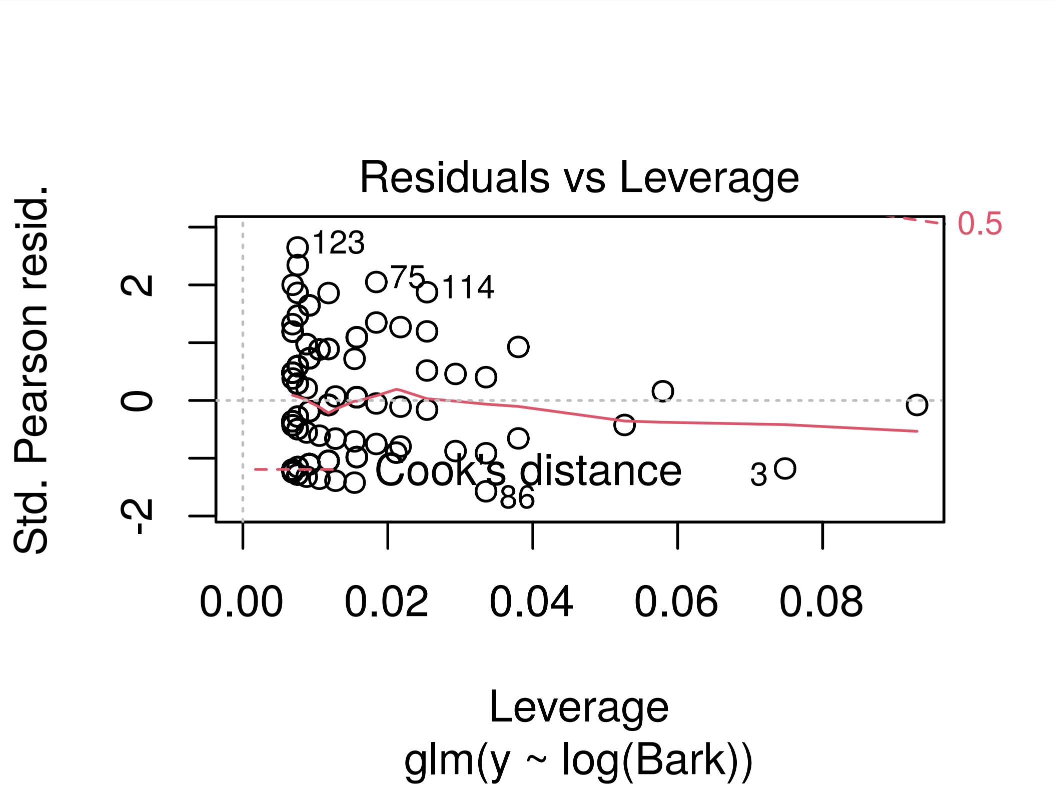


Why the "line" pattern?

- Looks "null" enough (the lines are due to the discrete count nature of the data- these cases share the same response but have different bark values)

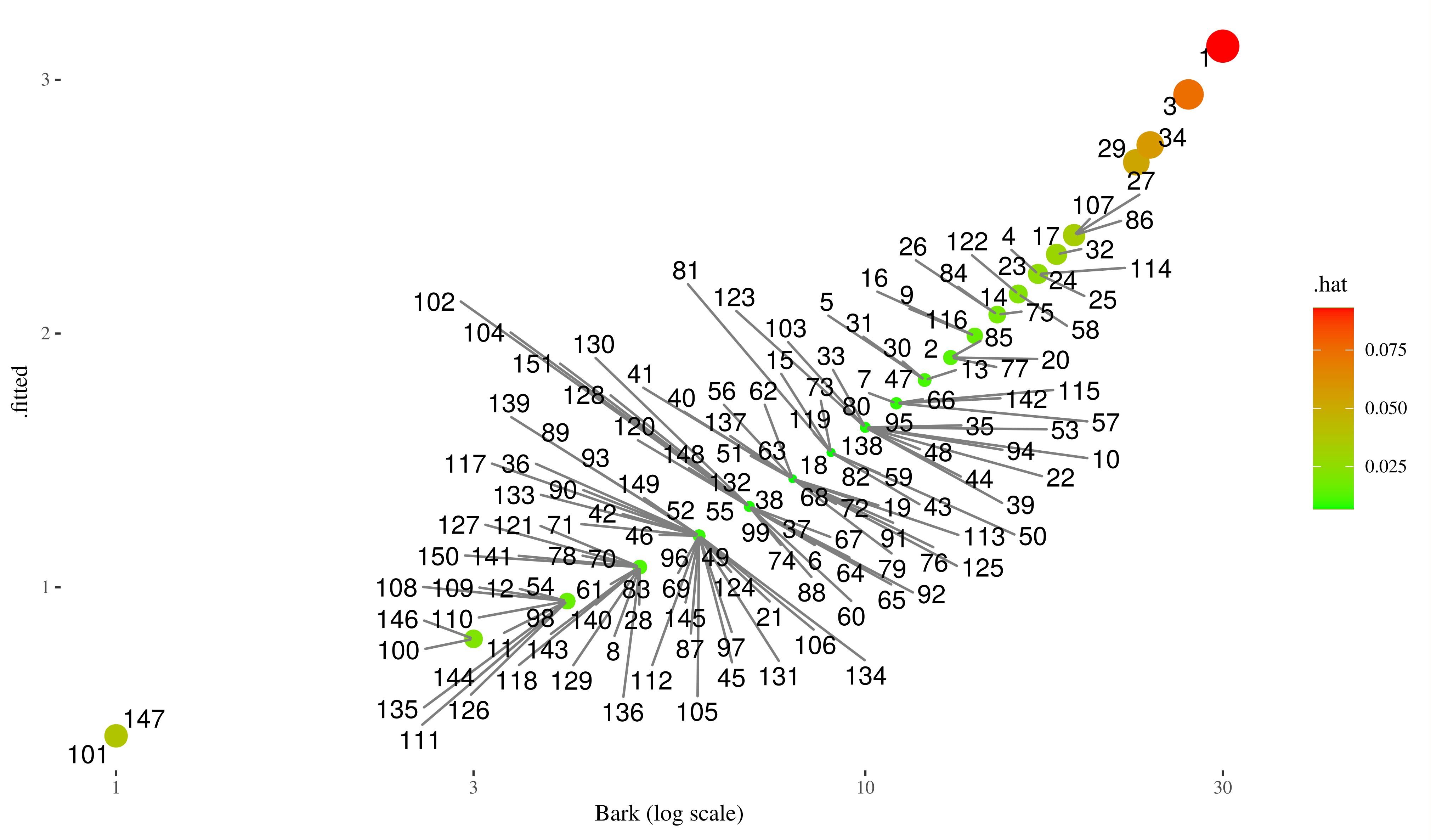
Australian Possums



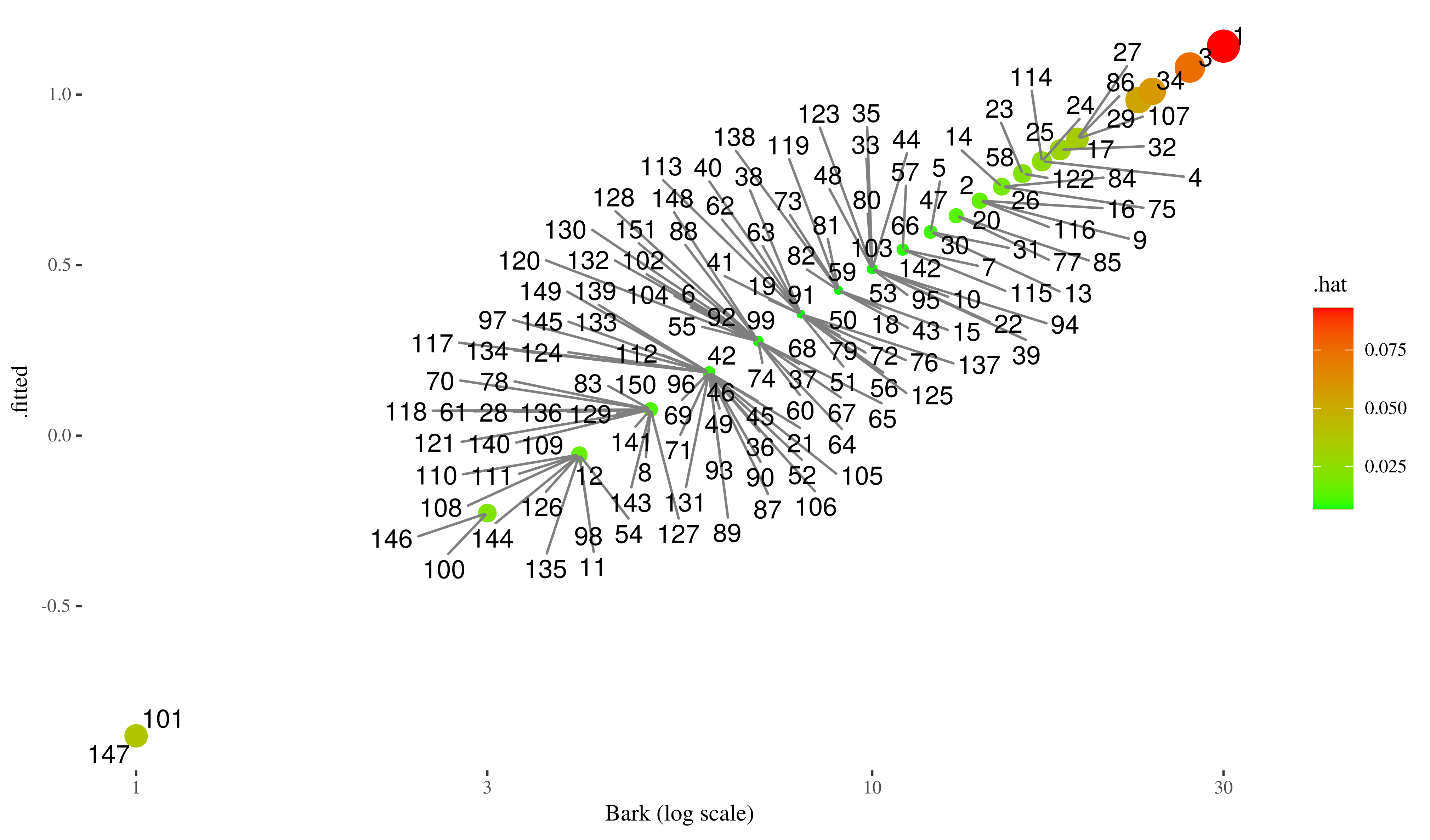


Highest Cook's D: Case 3 has little effect on the model (fit with and without)

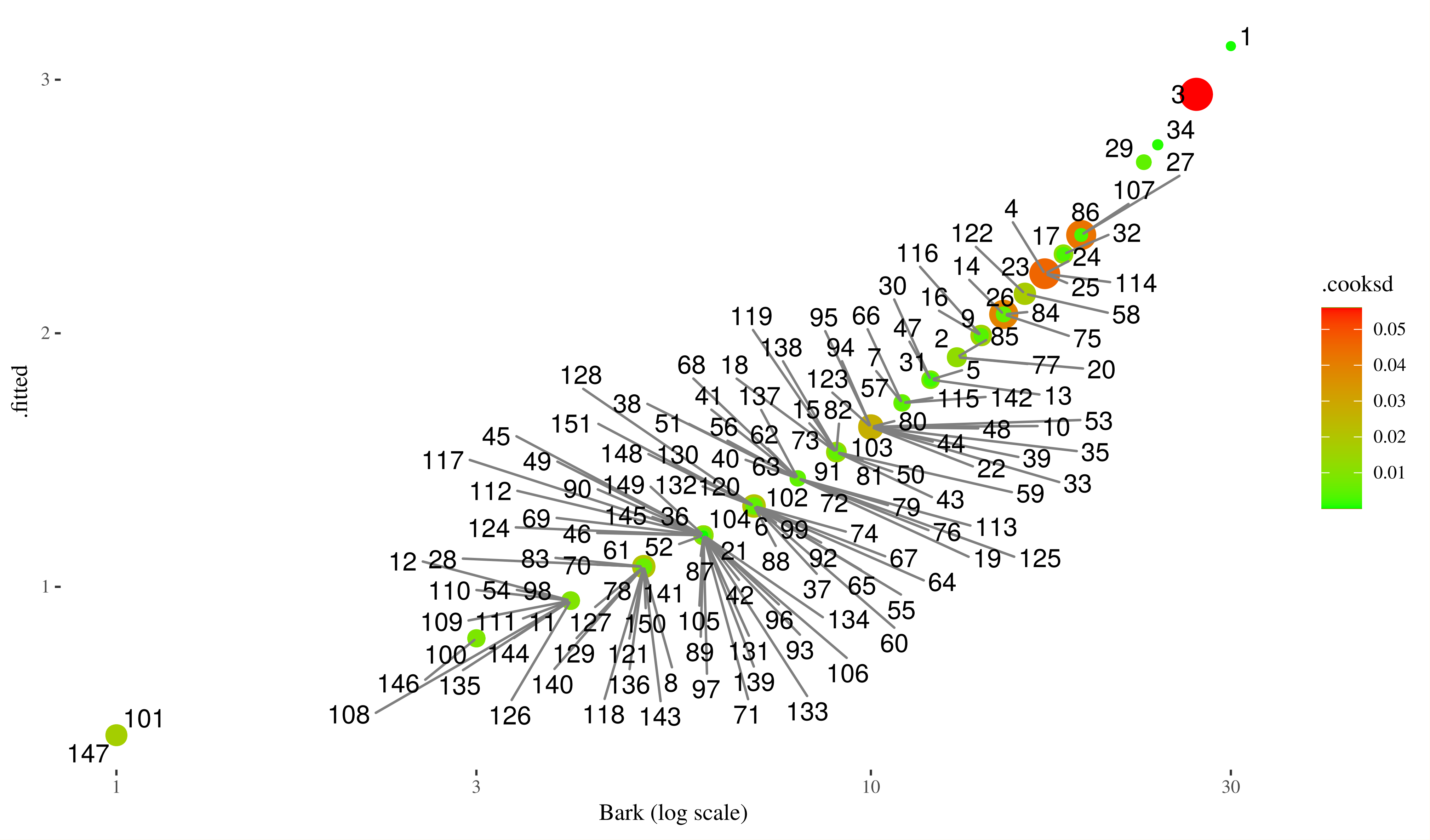
Original fit



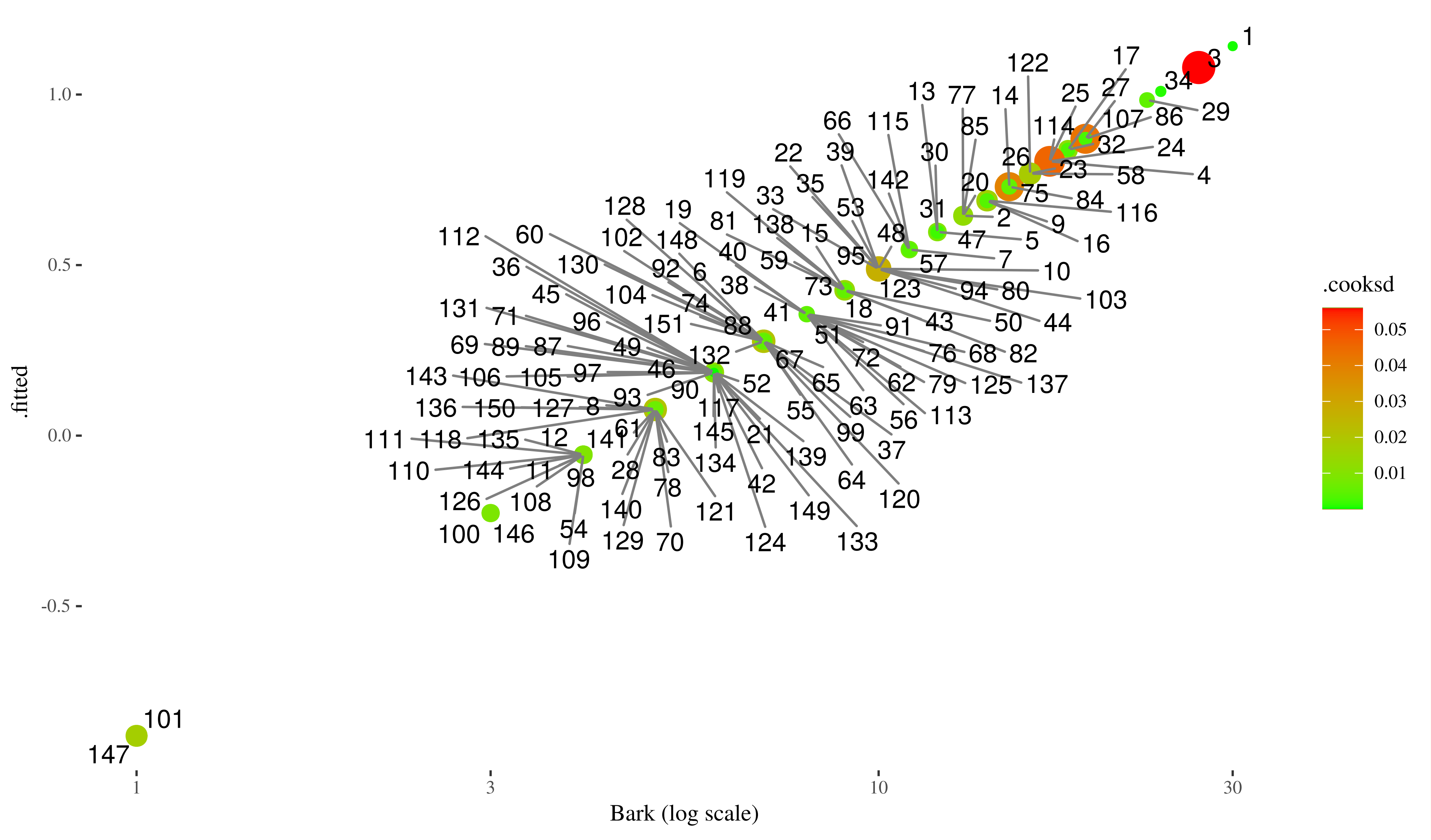
Logged fit


.fitted
1.0
0.5
0.0
-0.5
1
3
10
30
Bark (log scale)
.hat
0.075
0.050
0.025
1
3
101
147

Original fit



Logged fit



.fitted
1.0
0.5
0.0
-0.5
Bark (log scale)
1
3
10
30
.cooksd
0.05
0.04
0.03
0.02
0.01
1
17 3
122
34
27
13 77 14 25
29
107 86
66 115
85 114
32
24
22 39 30
26
23
58 4
119 142 20
35
75
33
84
128 19 53 31
116
2
81 9
138
112 60 102 148 15 48
47 5
16
40
95
59 7
130 57
36 92 6
10
38 73 123
94 80
45 104
103
74
131 18
41
71
96 43 50 44
151 88
91
51
69 89 87 49
132 76 68 82
72
67
46
97
143 106 105
65 62 79 125 137
52
90
93
63
136 150 127 8 117
55
113
56
61
145 21 99
111 118 135 12 141
37
83 139
134
110 144 11 28
98
64
42
78
120
126 108
140
121 149
54
100 146
129 70 124 133
109
101
147

Assessing Poisson model assumptions

Log-mean linearity:

- plot of log-response against quantitative predictors
- plot of residuals against quantitative predictors

Cases are independent:

- understanding of how the data was collected

Assessing Poisson model assumptions

counts of events Y_i has a **Poisson distribution** with mean and variance $\mu_{y|x}$

- Check residuals, should have equal scatter and spread around the 0line given any x value.
- Check goodness-of-fit test, compare sample means and variances of similar groups

Assessing Poisson model assumptions

When might your response count NOT follow a Poisson distribution?

The events do NOT occur independently

- could be clustering of "successes" in time or space
- clustering of event occurrences induces more variation in our responses than our Poisson model assumes.

Bad mean function

- Missing explanatory variables
- Incorrect mean function form (missing transformations, interactions, etc)

These issues can induce overdispersion, or extra-Poisson variation, in your response, resulting in SEs that are too small.

Goodness-of-fit test

$$\begin{aligned} H_0 &: \text{ Poisson model} \\ H_A &: \text{ saturated model} \end{aligned}$$

- The test statistic is the model's deviance G^2

$$G^2 = 2\left[\ln L\left(y_i\right) - \ln L(\hat{\mu}(X))\right]$$

- When $\hat{\mu}_i$'s are large (> 5), p-value is approximately

$$\text{p-value } = 1 - P\left(\chi^2 > G^2\right) = 1 - pchisq\left(G^2, df = n - (p+1)\right)$$

Australian Possums

```
summary(pos_glm)

Call:
glm(formula = y ~ log(Bark), family = poisson, data = possums)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.18523  -1.26246  -0.07764   0.55078   2.11368  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -0.8801     0.3027  -2.907  0.00365 ** 
log(Bark)     0.5945     0.1335   4.453 8.45e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 187.49  on 150  degrees of freedom
Residual deviance: 167.51  on 149  degrees of freedom
AIC: 452.31

Number of Fisher Scoring iterations: 5
```

Australian Possums

GOF test for the possums model (only using log bark):

```
1 - pchisq(167.51, df = 149)
[1] 0.1425203
```

Can we trust this test?

No! The chi-square model won't be a good approximation for the distribution of G^2 when data counts and model mean counts are small

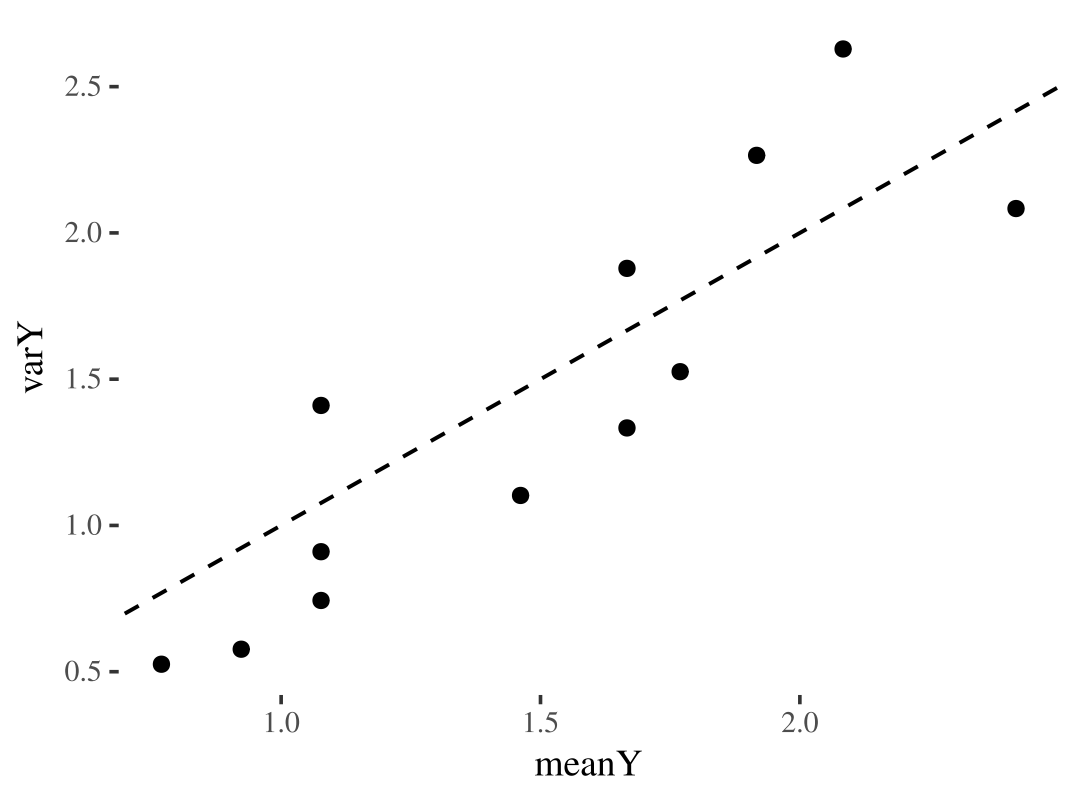
```
summary(possums_aug$.fitted) # all < 5
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.4147  1.2033  1.4277  1.4768  1.6302  3.1325 
```

GOF alternative visualization

```
pos_byBark <- possums %>%
 mutate(Bark_grp = ntile(Bark, n=12)) %>%
 group_by(Bark_grp) %>%
 summarize(meanY = mean(y), varY = var(y))

ggplot(pos_byBark, aes(x = meanY, y = varY)) +
 geom_point() +
geom_abline(intercept = 0, slope = 1, linetype = 2)
```



Quasi-Poisson model

What if we reject our GOF test or find visual evidence of extra-Poisson variation?

- Quasi-Poisson model

$$
\begin{aligned}
Y_i \mid x_i &\sim \text{Poisson}(\mu_i) \\
E\left(Y_i \mid x_i\right) &= \mu_i \\
V\left(Y_i \mid x_i\right) &= \psi\mu_i
\end{aligned}
$$

- A quasi-poisson model and drop-in-deviance:

```
glm(y ~ x1 + x2, family = quasipoisson, data = mydata)

anova(red_quasi, full_quasi, test = "F")
```

Estimating the dispersion parameter ψ

For a GLM, the dispersion parameter ψ ("psi") is estimated from the deviance G^2 from the regular GLM:

$$\hat{\psi} = \frac{G^2}{n - (p + 1)}$$

- $\hat{\psi} > 1$: overdispersion (responses are more variable than expected)
- $\hat{\psi} < 1$: underdispersion (responses are less variable than expected)

Quasi-Poisson model

- Conduct "z"-inference (Wald tests/CI) using SEs equal to $SE_{\text{quasi}}\left(\hat{\beta}_i\right)$
- Compare quasi-poisson models using a F-test stat equal to

$$F = \frac{\left(G^2_{\text{reduced}} - G^2_{\text{full}}\right)/(\#\text{ terms tested })}{\hat{\psi}}$$

using an F-distribution with degrees of freedom equal to the number of terms tested and $n-(p+1)$. G^2 is the model deviance from fitting the usual Poisson model for two competing models.