Assumptions

Ali Thursland and Mary Helen Wood

12/10/2018

== Setup ==

== Read in Data ==

Separate training and testing

```r
#80% of the sample size
smp_size <- floor(0.80 * nrow(airbnb))

#set the seed to make your partition reproducible
set.seed(123456)
train_ind <- sample(seq_len(nrow(airbnb)), size = smp_size)

train.airbnb <- airbnb[train_ind, ]
test.airbnb <- airbnb[-train_ind, ]
```

== Final Model ==

```r
stepwise.interactions.model <- lm(logprice ~ cleaning_fee * accommodates + availability_30 * minimum_ni
+ host_is_superhost
+ room_type
+ accommodates
+ cleaning_fee
+ minimum_nights
+ availability_30
+ log.reviews_per_month
+ cancellation_policy, data=train.airbnb)

kable(tidy(stepwise.interactions.model), format="markdown", digits = 4)
```

term	estimate	std.error	statistic	p.value
(Intercept)	4.3820	0.0482	90.8472	0.0000
cleaning_fee	0.0038	0.0005	7.8207	0.0000
accommodates	0.0982	0.0072	13.6651	0.0000
availability_30	0.0030	0.0015	1.9919	0.0466
minimum_nights	-0.0189	0.0020	-9.5745	0.0000
host_is_superhostt	0.0590	0.0240	2.4610	0.0140
room_typePrivate room	-0.2604	0.0272	-9.5797	0.0000
room_typeShared room	-1.1917	0.1544	-7.7199	0.0000
log.reviews_per_month	-0.2061	0.0185	-11.1291	0.0000
cancellation_policymoderate	0.0730	0.0280	2.6076	0.0092
cancellation_policystrict_14_with_grace_period	0.1359	0.0306	4.4397	0.0000

term	estimate	std.error	statistic	p.value
cancellation_policysuper_strict_30	0.0727	0.1023	0.7110	0.4772
cancellation_policysuper_strict_60	0.7443	0.0841	8.8522	0.0000
cleaning_fee:accommodates	-0.0002	0.0001	-3.6396	0.0003
availability_30:minimum_nights	0.0003	0.0001	3.2409	0.0012

```r
kable(glance(stepwise.interactions.model))
```

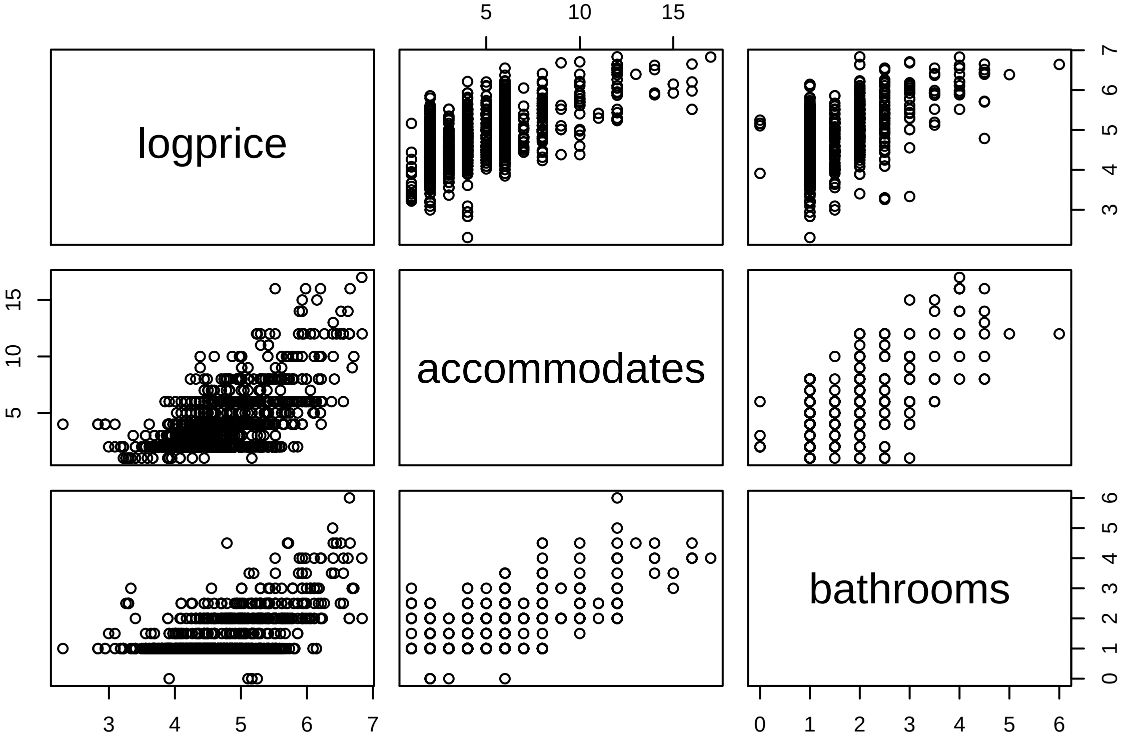
	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.r
value	0.5742094	0.5702486	0.4000556	144.9715	0	15	-756.6985	1545.397	1630.621	240.8669	
“ ‘											

== Assumptions ==

There are 5 assumptions for multiple linear regression: 1. Linearity 2. Constant variance 3. Normality 4. Independence

Additionally, we must avoid outliers/multicollinearity in our final model.

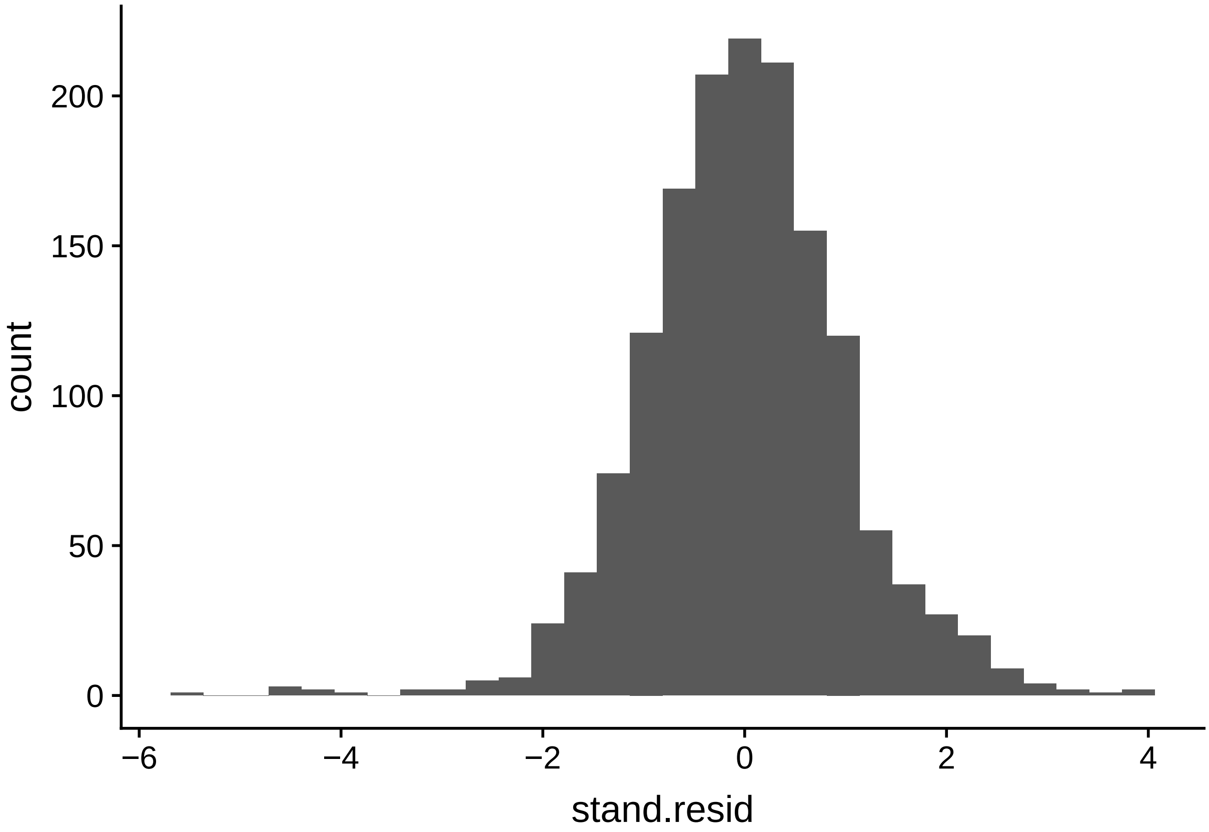
```r
pairs(logprice ~ accommodates + bathrooms, data = train.airbnb)
```



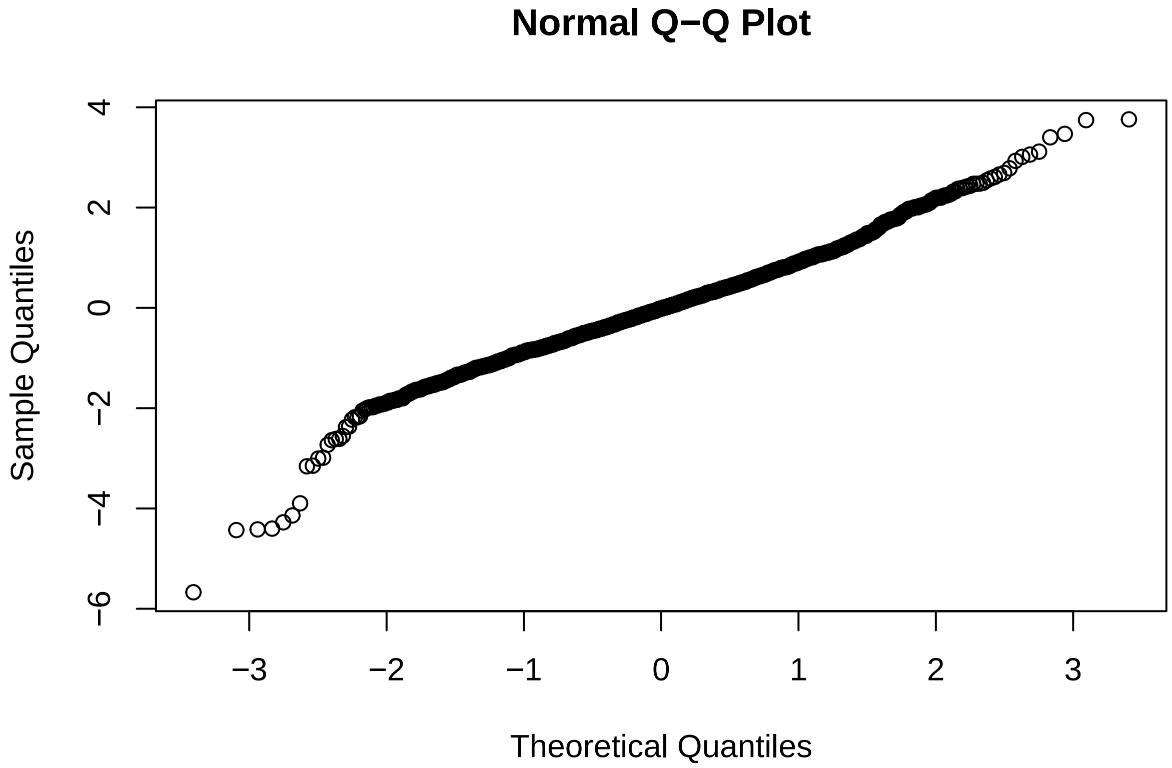
```r
train.airbnb <- train.airbnb %>% mutate(stand.resid = rstandard(stepwise.interactions.model),
                                        pred = predict(stepwise.interactions.model))
```

```r
ggplot(data = train.airbnb, aes(x=stand.resid)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

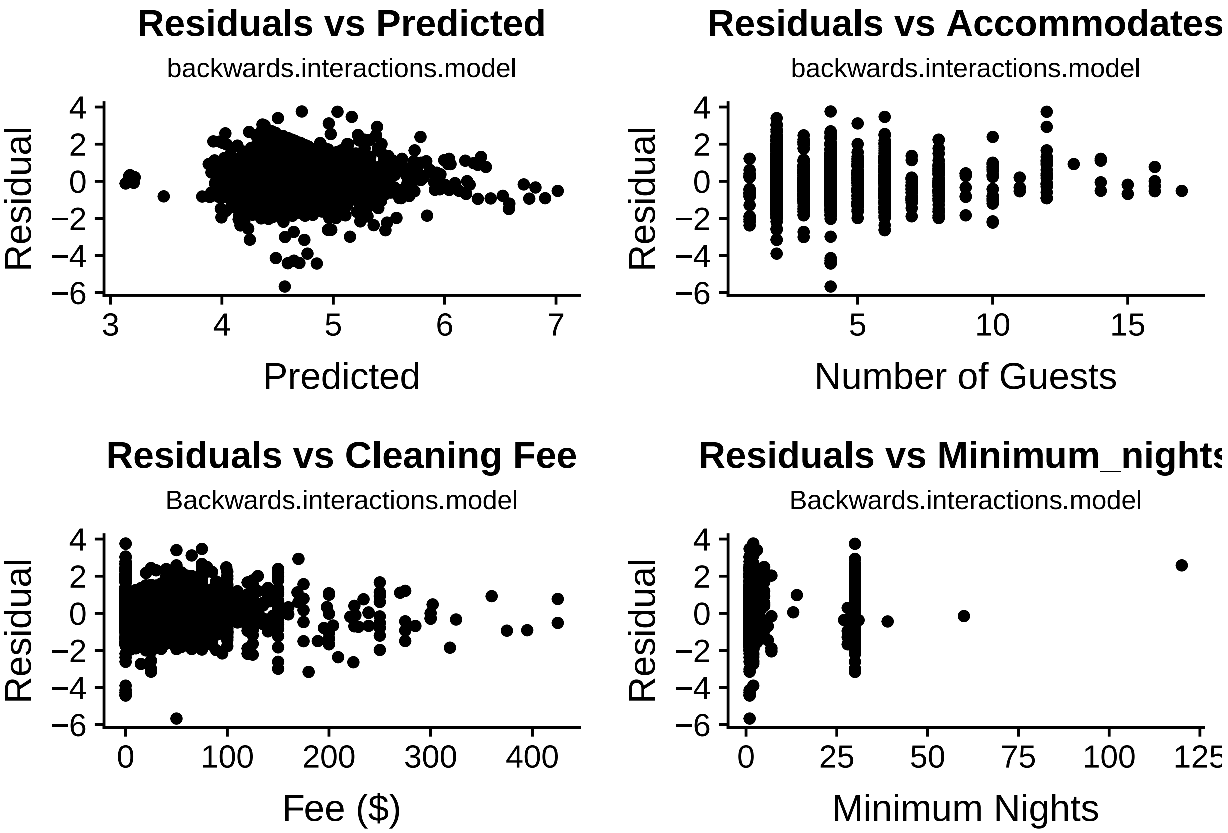


```r
qqnorm(train.airbnb$stand.resid)
```

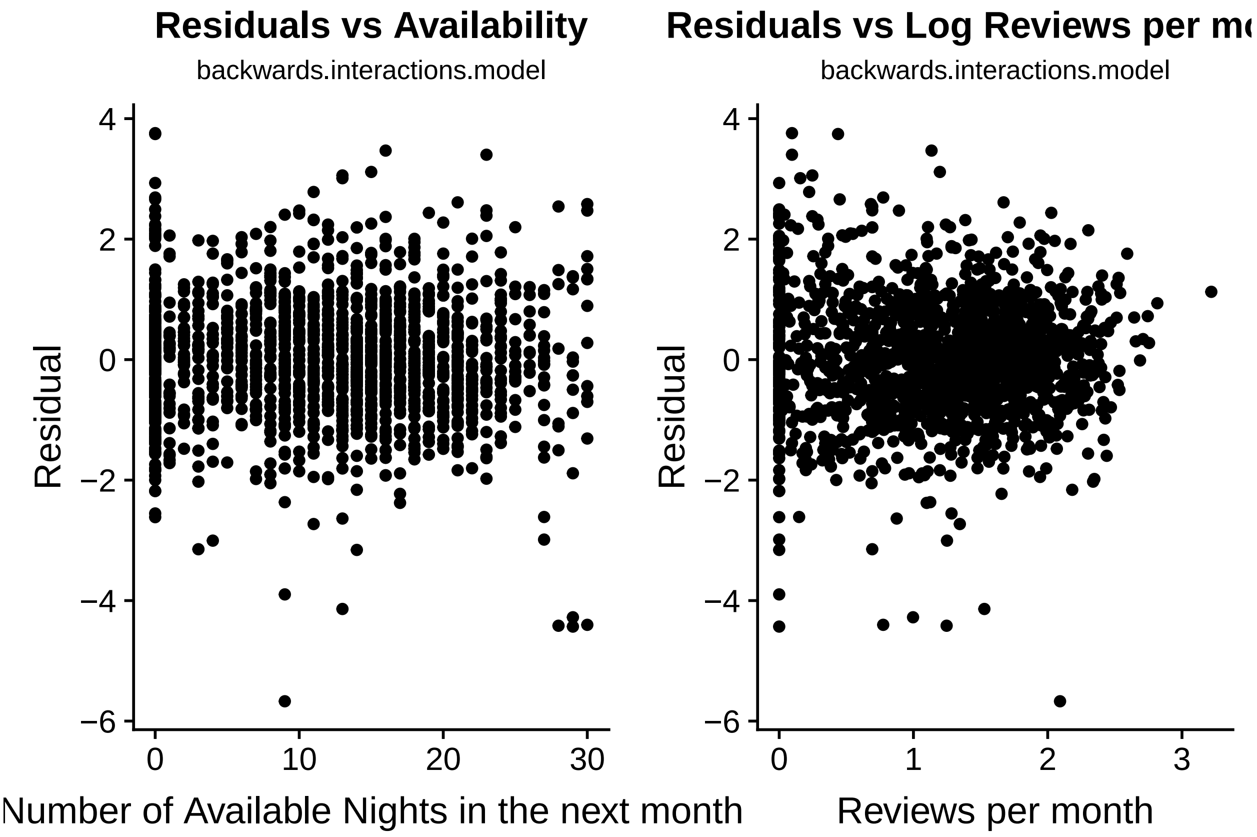


```
p1 <- ggplot(data = train.airbnb, aes(x=pred, y=stand.resid)) + geom_point() +
  labs(x="Predicted", y="Residual", title="Residuals vs Predicted",
subtitle=("backwards.interactions.model"))+
theme(plot.title = element_text(hjust = 0.5,size=14),
plot.subtitle=element_text(hjust=0.5,size=10))


p2 <- ggplot(data = train.airbnb, aes(x=accommodates, y=stand.resid)) + geom_point() +
  labs(x="Number of Guests", y="Residual", title="Residuals vs Accommodates",
subtitle=("backwards.interactions.model"))+
theme(plot.title = element_text(hjust = 0.5,size=14),
plot.subtitle=element_text(hjust=0.5,size=10))

p3 <- ggplot(data = train.airbnb, aes(x= cleaning_fee, y=stand.resid)) + geom_point() +
  labs(x="Fee ($)", y="Residual", title="Residuals vs Cleaning Fee",
subtitle=("Backwards.interactions.model"))+
theme(plot.title = element_text(hjust = 0.5,size=14),
plot.subtitle=element_text(hjust=0.5,size=10))

p4 <-ggplot(data = train.airbnb, aes(x=minimum_nights, y=stand.resid)) + geom_point() +
  labs(x="Minimum Nights", y="Residual", title="Residuals vs Minimum_nights",
subtitle=("Backwards.interactions.model"))+
theme(plot.title = element_text(hjust = 0.5,size=14),
plot.subtitle=element_text(hjust=0.5,size=10))

p5 <- ggplot(data = train.airbnb, aes(x=availability_30, y=stand.resid)) + geom_point() +
  labs(x="Number of Available Nights in the next month", y="Residual", title="Residuals vs Availability
subtitle=("backwards.interactions.model"))+
theme(plot.title = element_text(hjust = 0.5,size=14),
plot.subtitle=element_text(hjust=0.5,size=10))

p6 <- ggplot(data = train.airbnb, aes(x=log.reviews_per_month, y=stand.resid)) + geom_point() +
  labs(x="Reviews per month", y="Residual", title="Residuals vs Log Reviews per month",
subtitle=("backwards.interactions.model"))+
theme(plot.title = element_text(hjust = 0.5,size=14),
plot.subtitle=element_text(hjust=0.5,size=10))
```

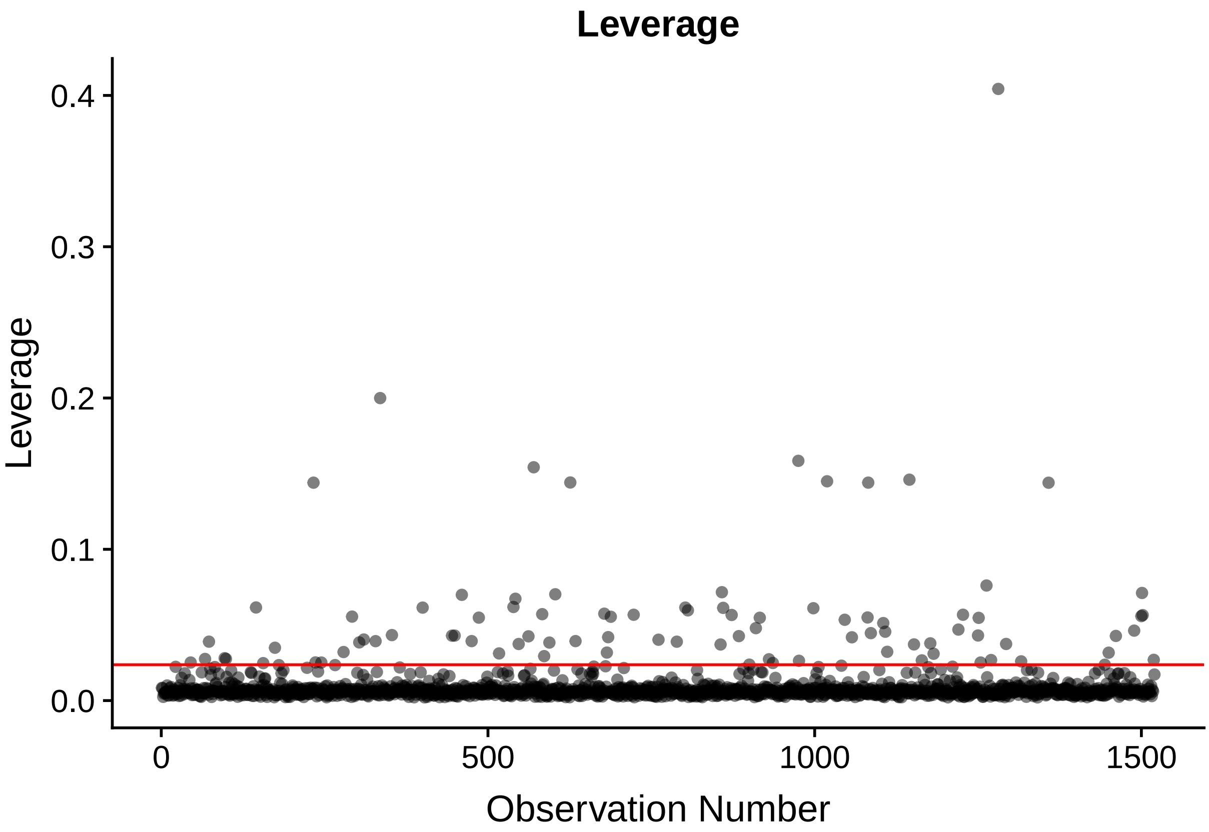
```
plot_grid(p1,p2,p3,p4)
```



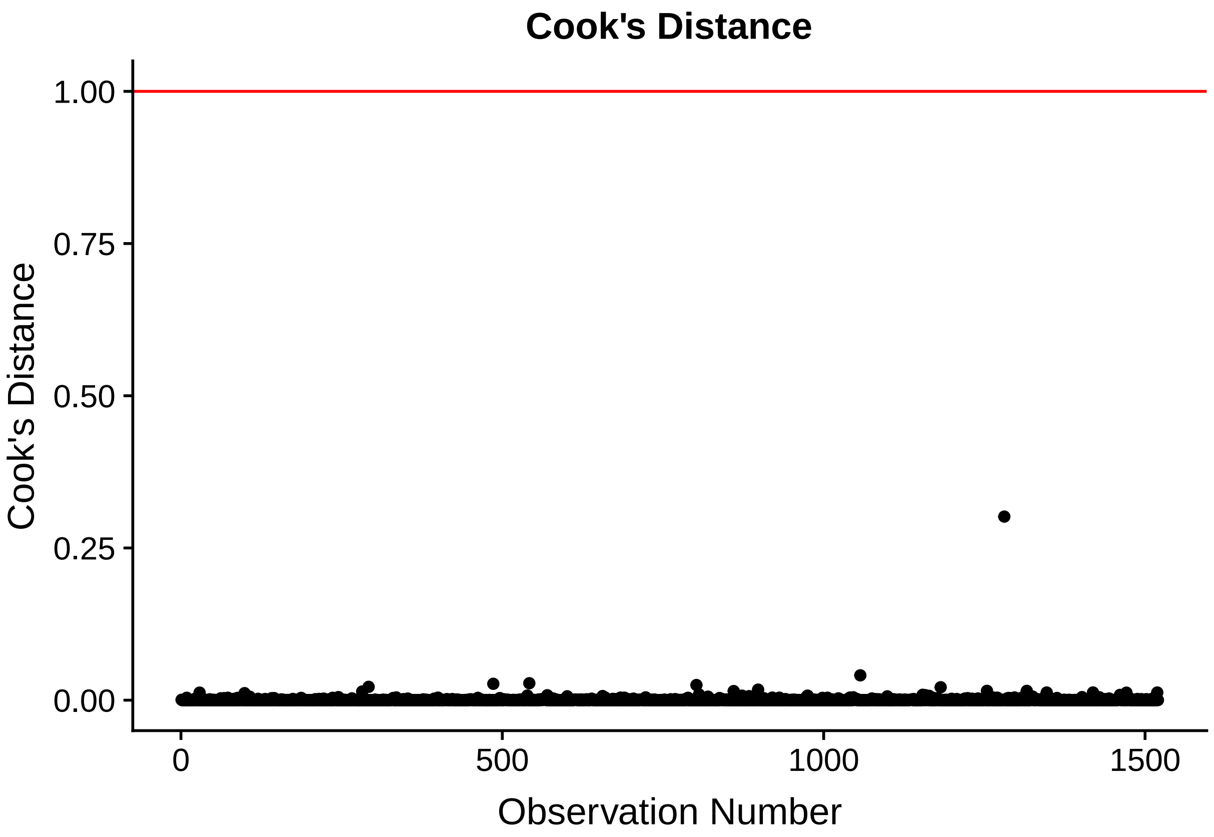
```
plot_grid(p5,p6)
```



```r
train.airbnb <- train.airbnb %>%
  mutate(leverage = hatvalues(stepwise.interactions.model),
         cooks = cooks.distance(stepwise.interactions.model),
         obs.num = row_number())

ggplot(data=train.airbnb, aes(x=obs.num,y=leverage)) +
  geom_point(alpha=0.5) +
  geom_hline(yintercept=36/1520,color="red")+
  labs(x="Observation Number",y="Leverage",title="Leverage")
```



```r
ggplot(data=train.airbnb, aes(x=obs.num,y=cooks)) +
  geom_point() +
  geom_hline(yintercept=1,color="red")+
  labs(x="Observation Number",y="Cook's Distance",title="Cook's Distance")
```



```
tidy(vif(stepwise.interactions.model))
```

```
## Warning: 'tidy.matrix' is deprecated.
## See help("Deprecated")
```

```
## # A tibble: 10 x 4
##    .rownames                       GVIF    Df GVIF..1..2.Df..
##    <chr>                          <dbl> <dbl>           <dbl>
##  1 cleaning_fee                    6.38     1            2.53
##  2 accommodates                    2.75     1            1.66
##  3 availability_30                 1.35     1            1.16
##  4 minimum_nights                  2.53     1            1.59
##  5 host_is_superhost               1.30     1            1.14
##  6 room_type                       1.46     2            1.10
##  7 log.reviews_per_month           1.49     1            1.22
##  8 cancellation_policy             1.55     4            1.06
##  9 cleaning_fee:accommodates       7.58     1            2.75
## 10 availability_30:minimum_nights  2.51     1            1.59
```