Spurious lag climate growth correlations

Question

- Does negative size x growth dependence create spurious climate lag effects in models of plant performance?

Analysis:

- Simulate growth of n individuals over t years using a log Gompertz model.
- Growth is size dependent, meaning large plants tend to grow proportionally slower. This prevents plants from growing indefinitely.
- Growth is sensitive to fluctuating annual environment symbolized by E.
- Individual variation in growth parameters means some plants may reach a larger equilibrium size than others.

```
# Parmaeters:
set.seed(1)
t <- 30
n <- 500
a <- 0.5 # Intrinsic growth rate
b <- 0.5 # Size limitation
E <- rnorm(t) # Annual environental variation
beta <- 0.25 # Environment effect
init_size <- 0.1 # Initial size
sig <- 0.3 # process noise

X <- matrix(NA, t, n)
X[1,] <- init_size

for ( i in 2:t ){
    mu <- a + b*X[i-1, ] + beta*E[i-1]
    X[i, ] <- rnorm(n,  mu, sig)
}

dat <-
  data.frame(t = 1:t , size = X, E = E) %>%
  gather( id, size, starts_with('size')) %>%
  mutate( id = as.numeric(factor( id )))

E_plot <-
  dat %>%
  distinct( t, E ) %>%
  ggplot( aes( x = t, y = E)) +
  geom_line()  +
  theme_bw() +
  theme(axis.title.x = element_blank(),
        axis.text.x = element_blank())

x_plot <-
  dat %>%
```

```
  filter( id < 8) %>%
  ggplot( aes( x = t, y = size, group = id, color = factor( id ) )) +
  geom_line( alpha = 0.9)  +
  theme_bw()  +
  scale_y_continuous(name = 'Size (log)' ) +
  scale_x_continuous(name = 'time') +
  scale_color_discrete(guide = 'none')


plot_grid(E_plot, x_plot, align = 'v' , ncol = 1)
```

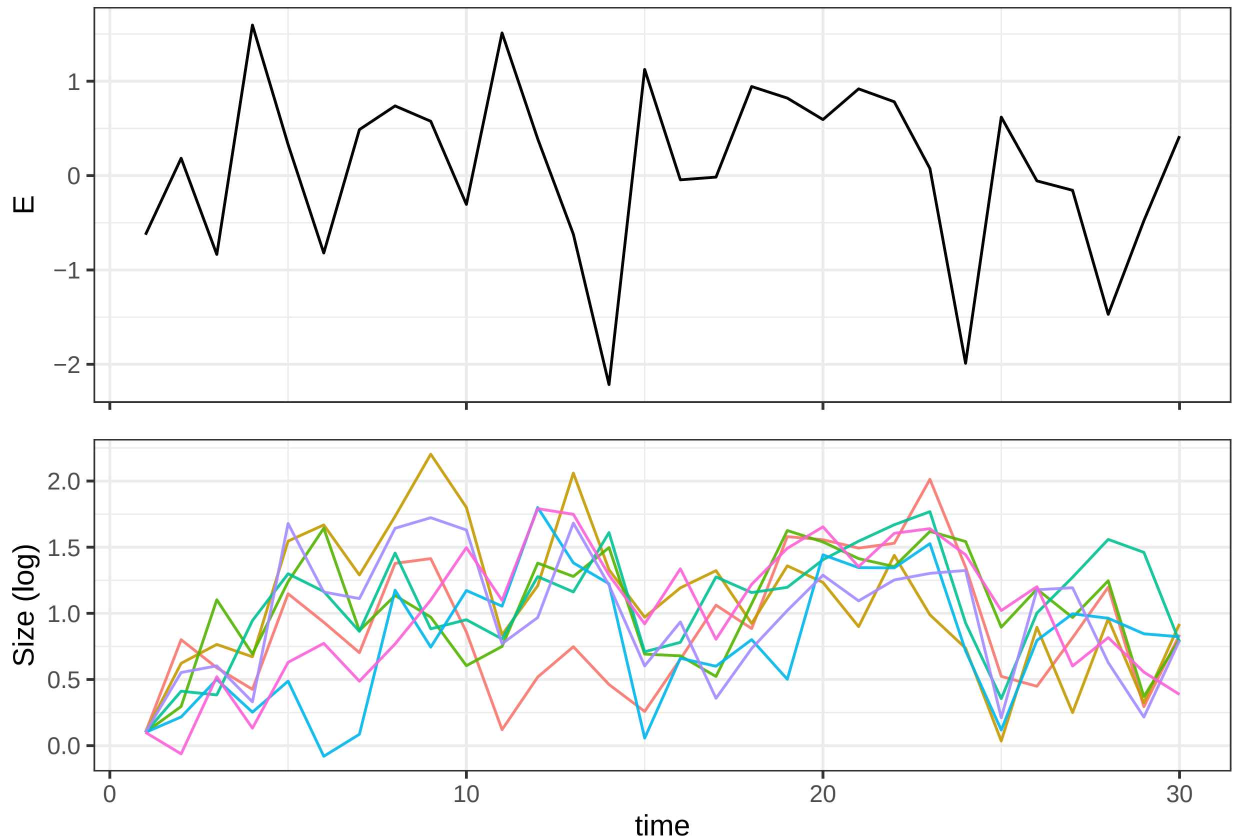


Figure 1. Simulated environmental variation (upper plot) and simulated size of a eight individual plants (lower plot) over 30 years.

In the simulation, growth between year t and t + 1 is only sensitive to the environment in year t. No other lagged years of climate contribute to growth.

Now we can check for correlations between various lags in the time series.

```
dat <-
  dat %>%
  group_by( id ) %>%
  arrange( id, t ) %>%
  mutate(size0 = lag(size), lag1 = lag( E ), lag2 = lag( E, 2 ), lag3 = lag(E, 3), lag4 = lag(E, 4)) %>%
  mutate( log_growth = size - size0) %>%
  mutate( rll_E = rollmean( lag( E, 2), k = 4, align = 'right', fill = NA))

dat_long <-
```

```
  dat %>%
  ungroup() %>%
  filter( complete.cases(.)) %>%
  gather( lag, Env, starts_with('lag'), rll_E)

lag_cors <-
  dat_long %>%
  group_by( lag ) %>%
  summarise( r = round( cor( log_growth, Env ), 2)) %>%
  mutate( label = paste0( 'r = ', r), pos_x = 0.5, pos_y = max(dat$log_growth, na.rm = T )*0.95)

subsamp <- dat_long %>% group_by( t ) %>% sample_frac(0.1, replace = F) # subsample to draw fewer point

dat_long %>%
  ggplot( aes( x = Env, y = log_growth )) +
  geom_point( data = subsamp  , aes( x = Env, y= log_growth ), alpha = 0.1) +
  geom_smooth( method = 'lm', se = F) +
  geom_text( data = lag_cors, aes( label = label, x = pos_x, y = pos_y), hjust = 1) +
  facet_wrap( ~ lag, nrow = 1, scales = 'free_x')
```

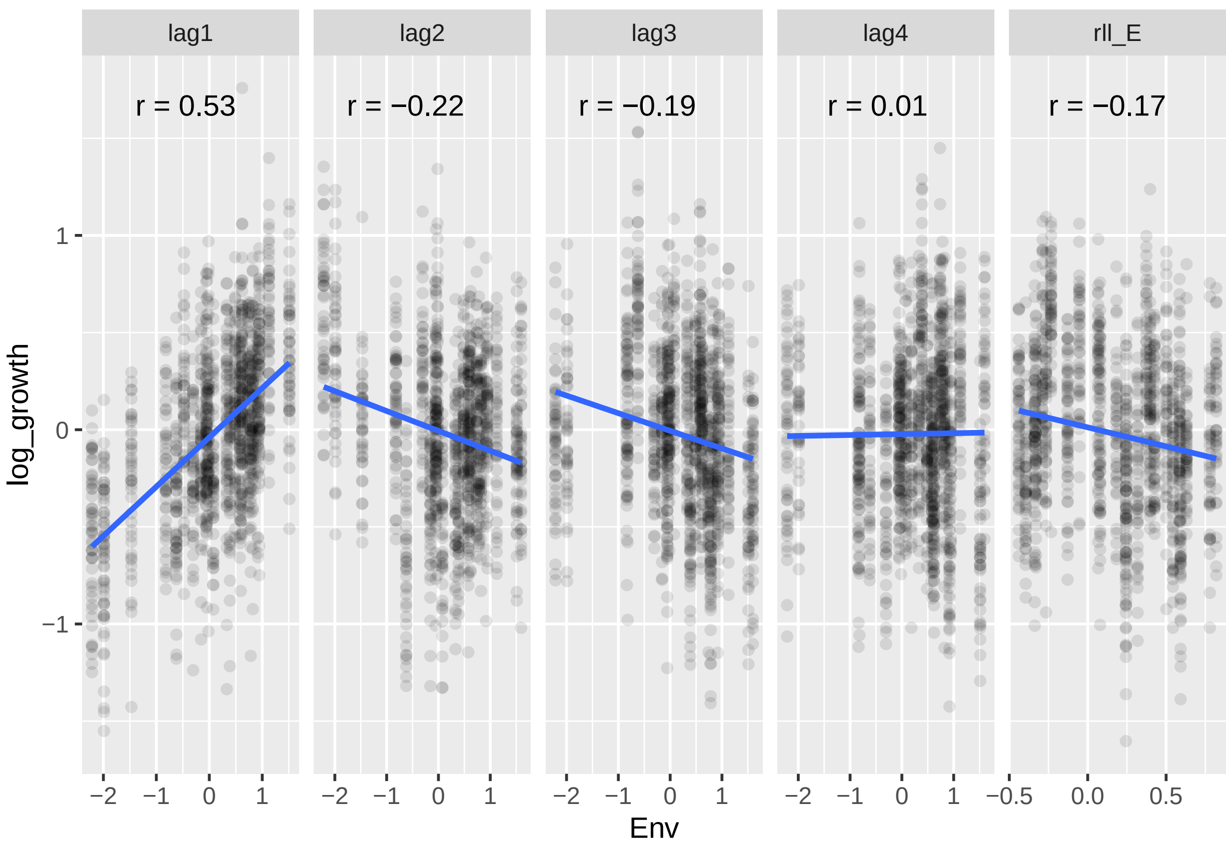


Figure 2. Correlation between log growth rate of each individual each year and environmental conditions at 4 different lags. "rll_E" is rolling mean of Env for four years prior to transition.

Growth between year t and t + 1 is positively correlated with E in year t (identified as "lag1" in the figure). It is negatively correlated with E lagged over 1 to 3 years ("lag2" through "lag4"). It is also negatively correlated with a rolling mean of E from year t - 4 to t - 1.

Now I attempt to fit a linear model to the data to describe size dependence and environmental effects.

```
m1 <- lm( data = dat, size ~ size0 + lag1 + lag2 + lag3 + lag4 + rll_E )

summary(m1)
```

```
##
## Call:
## lm(formula = size ~ size0 + lag1 + lag2 + lag3 + lag4 + rll_E,
##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29078 -0.20073 -0.00237  0.20181  1.11889
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.504043   0.008260  61.024   <2e-16 ***
## size0        0.493464   0.007745  63.710   <2e-16 ***
## lag1         0.251887   0.003043  82.771   <2e-16 ***
## lag2         0.006360   0.004285   1.484   0.1378
## lag3         0.004495   0.004015   1.119   0.2630
## lag4         0.007564   0.003976   1.902   0.0572 .
## rll_E       -0.016343   0.012123  -1.348   0.1777
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2994 on 12493 degrees of freedom
##   (2500 observations deleted due to missingness)
## Multiple R-squared:  0.5186, Adjusted R-squared:  0.5184
## F-statistic:  2243 on 6 and 12493 DF,  p-value: < 2.2e-16
```

The model estimates the parameters of the simulation accurately. Despite the negative correlation between lagged climate and growth observed in the figure above, there are no strong lagged climate effects in the model.

The above simulation and model fitting shows that a positive climate effect in one year can produce spurious negative correlations between growth and lagged climate. However, accounting for size dependence in the growth model fixes this problem.

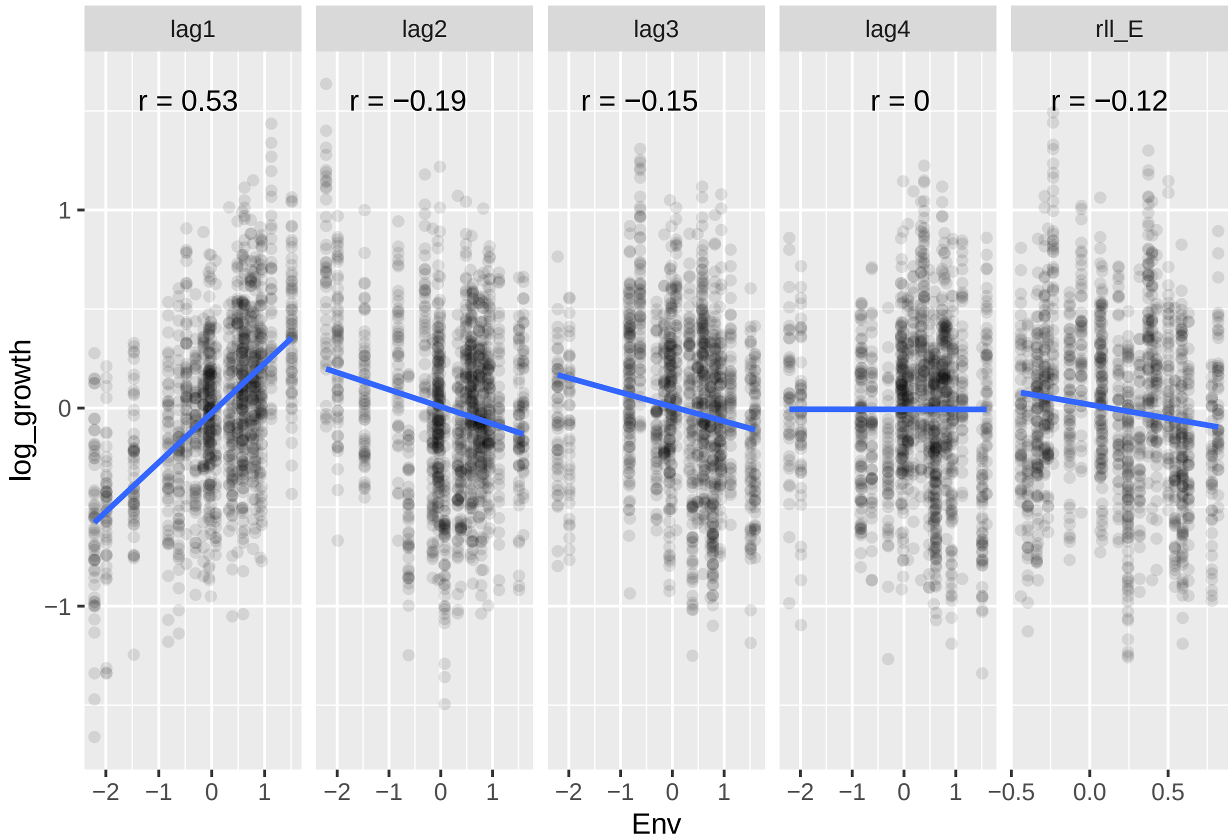
Including individual variation in growth

The above simulation and model fitting process assumed all individual plants had the same underlying growth parameters (a and b above). However, its reasonable to expect that these processes vary by plant to some degree. We can simulate this by introducing noise into growth parameters a and/or b.

```
# Other parematers as above

b <- runif(n,  0.2 , 0.9 ) # introduce noise in b
X <- matrix(NA, t, n)
X[1,] <- init_size

for ( i in 2:t ){
    mu <- a + b*X[i-1, ] + beta*E[i-1]
    X[i, ] <- rnorm(n,  mu, sig)
```

```
}

dat <-
  data.frame(t = 1:t , size = X, E = E) %>%
  gather( id, size, starts_with('size')) %>%
  mutate( id = as.numeric(factor( id )))

E_plot <-
  dat %>%
  distinct( t, E ) %>%
  ggplot( aes( x = t, y = E)) +
  geom_line()  +
  theme_bw() +
  theme(axis.title.x = element_blank(),
        axis.text.x = element_blank())

x_plot <-
  dat %>%
  filter( id < 8) %>%
  ggplot( aes( x = t, y = size, group = id, color = factor( id ) )) +
  geom_line( alpha = 0.9)  +
  theme_bw()  +
  scale_y_continuous(name = 'Size (log)' ) +
  scale_x_continuous(name = 'time') +
  scale_color_discrete(guide = 'none')

#plot_grid(E_plot, x_plot, align = 'v' , ncol = 1)

dat <-
  dat %>%
  group_by( id ) %>%
  arrange( id, t ) %>%
  mutate(size0 = lag(size), lag1 = lag( E ), lag2 = lag( E, 2 ), lag3 = lag(E, 3), lag4 = lag(E, 4)) %>
  mutate( log_growth = size - size0) %>%
  mutate( rll_E = rollmean( lag( E, 2), k = 4, align = 'right', fill = NA))

dat_long <-
  dat %>%
  ungroup() %>%
  filter( complete.cases(.)) %>%
  gather( lag, Env, starts_with('lag'), rll_E)

lag_cors <-
  dat_long %>%
  group_by( lag ) %>%
  summarise( r = round( cor( log_growth, Env ), 2)) %>%
  mutate( label = paste0( 'r = ', r), pos_x = 0.5, pos_y = max(dat$log_growth, na.rm = T )*0.95)


subsamp <- dat_long %>% group_by( t ) %>% sample_frac(0.1, replace = F) # subsample to draw fewer point

dat_long %>%
  ggplot( aes( x = Env, y = log_growth )) +
```

```r
  geom_point( data = subsamp  , aes( x = Env, y= log_growth ), alpha = 0.1) +
  geom_smooth( method = 'lm', se = F) +
  geom_text( data = lag_cors, aes( label = label, x = pos_x, y = pos_y), hjust = 1) +
  facet_wrap( ~ lag, nrow = 1, scales = 'free_x')
```




```r
m1 <- lm( data = dat, size ~ size0 + lag1 + lag2 + lag3 + lag4 + rll_E )

summary(m1)
```

```
##
## Call:
## lm(formula = size ~ size0 + lag1 + lag2 + lag3 + lag4 + rll_E,
##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.43933 -0.23804  0.00258  0.23927  1.37253
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.082536   0.005881  14.034  < 2e-16 ***
## size0        0.947577   0.003239 292.542  < 2e-16 ***
## lag1         0.248245   0.003561  69.710  < 2e-16 ***
## lag2        -0.090445   0.004655 -19.431  < 2e-16 ***
## lag3        -0.032794   0.004630  -7.083 1.49e-12 ***
## lag4        -0.018295   0.004646  -3.937 8.28e-05 ***
```

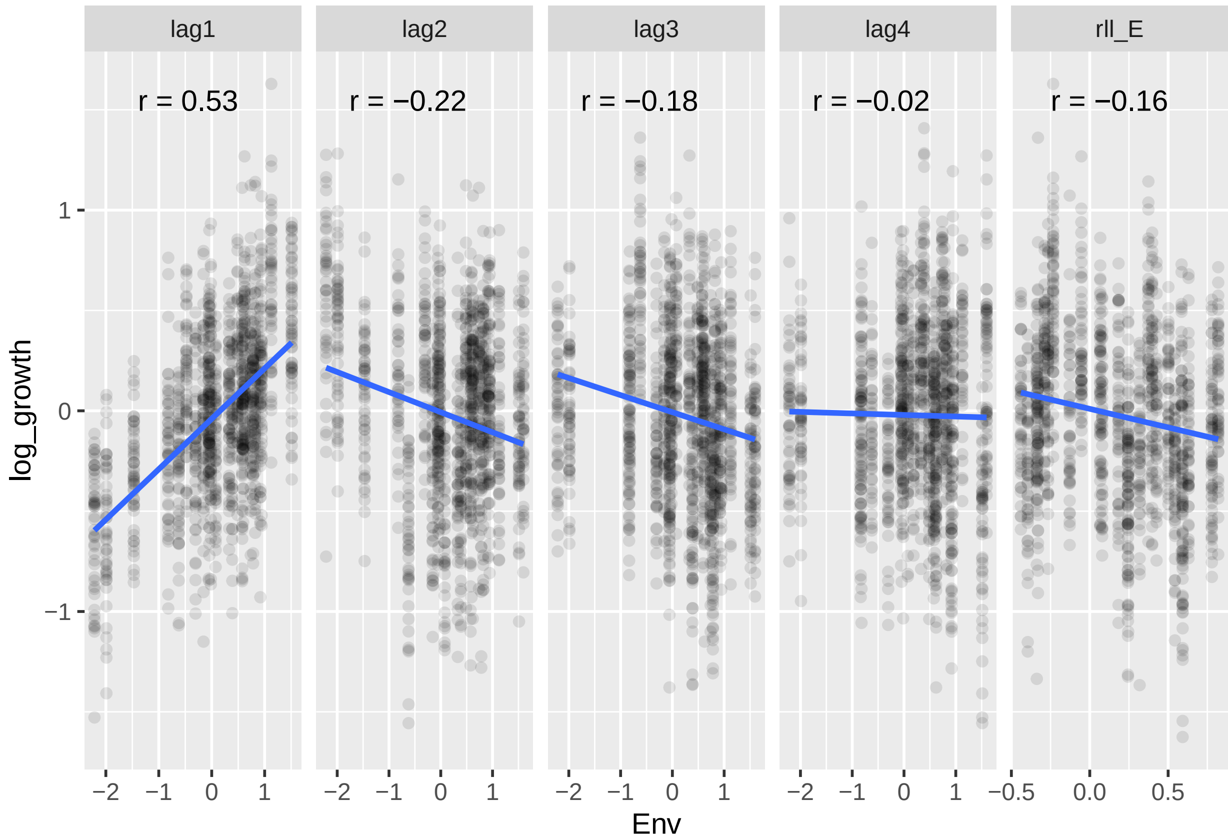
```
## rll_E       -0.012595   0.014143  -0.891    0.373
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3502 on 12493 degrees of freedom
##   (2500 observations deleted due to missingness)
## Multiple R-squared:  0.8796, Adjusted R-squared:  0.8795
## F-statistic: 1.521e+04 on 6 and 12493 DF,  p-value: < 2.2e-16
```

Because individuals vary in the parameter b fitting only a single parameter estimate for all individuals leads to a spurious negative climate effect showing up in the model.

We can repeat this as above except introducing random variation in a.

```
b <- 0.5
a <- rbeta(n, 2, 2) # introduce noise in a
X <- matrix(NA, t, n)
X[1,] <- init_size

for ( i in 2:t ){
    mu <- a + b*X[i-1, ] + beta*E[i-1]
    X[i, ] <- rnorm(n,  mu, sig)
}

dat <-
  data.frame(t = 1:t , size = X, E = E) %>%
  gather( id, size, starts_with('size')) %>%
  mutate( id = as.numeric(factor( id )))

dat <-
  dat %>%
  group_by( id ) %>%
  arrange( id, t ) %>%
  mutate(size0 = lag(size), lag1 = lag( E ), lag2 = lag( E, 2 ), lag3 = lag(E, 3), lag4 = lag(E, 4)) %>
  mutate( log_growth = size - size0) %>%
  mutate( rll_E = rollmean( lag( E, 2), k = 4, align = 'right', fill = NA))

dat_long <-
  dat %>%
  ungroup() %>%
  filter( complete.cases(.)) %>%
  gather( lag, Env, starts_with('lag'), rll_E)

lag_cors <-
  dat_long %>%
  group_by( lag ) %>%
  summarise( r = round( cor( log_growth, Env ), 2)) %>%
  mutate( label = paste0( 'r = ', r), pos_x = 0.5, pos_y = max(dat$log_growth, na.rm = T )*0.95)


subsamp <- dat_long %>% group_by( t ) %>% sample_frac(0.1, replace = F) # subsample to draw fewer point

dat_long %>%
  ggplot( aes( x = Env, y = log_growth )) +
```

```r
  geom_point( data = subsamp  , aes( x = Env, y= log_growth ), alpha = 0.1) +
  geom_smooth( method = 'lm', se = F) +
  geom_text( data = lag_cors, aes( label = label, x = pos_x, y = pos_y), hjust = 1) +
  facet_wrap( ~ lag, nrow = 1, scales = 'free_x')
```




```r
m1 <- lm( data = dat, size ~ size0 + lag1 + lag2 + lag3 + lag4 + rll_E )
summary(m1)
```

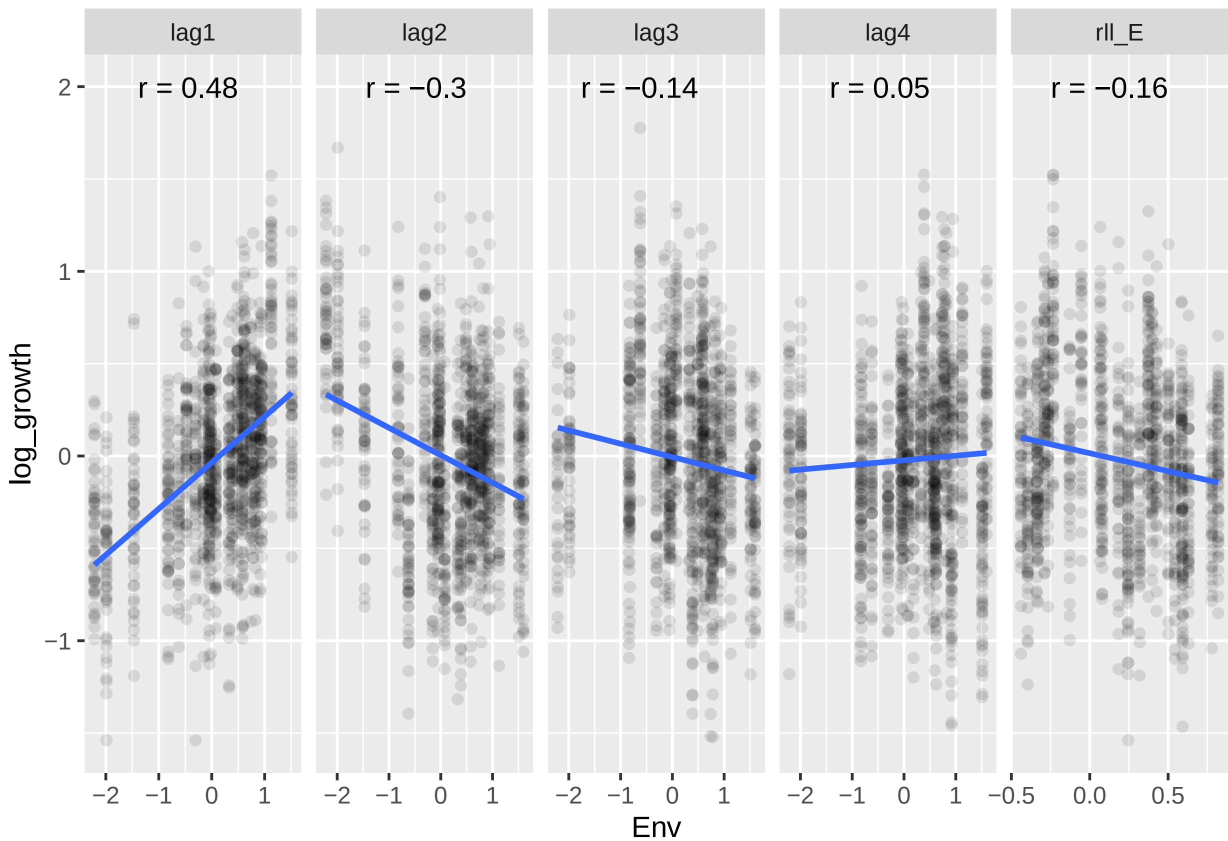
```
##
## Call:
## lm(formula = size ~ size0 + lag1 + lag2 + lag3 + lag4 + rll_E,
##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32737 -0.22506  0.00205  0.22811  1.14718
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.178202   0.006095  29.237  < 2e-16 ***
## size0        0.816799   0.005200 157.069  < 2e-16 ***
## lag1         0.248893   0.003415  72.889  < 2e-16 ***
## lag2        -0.072516   0.004582 -15.828  < 2e-16 ***
## lag3        -0.030576   0.004465  -6.847 7.87e-12 ***
## lag4        -0.021466   0.004461  -4.812 1.51e-06 ***
## rll_E       -0.016429   0.013560  -1.212    0.226
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3359 on 12493 degrees of freedom
##   (2500 observations deleted due to missingness)
## Multiple R-squared:  0.7185, Adjusted R-squared:  0.7183
## F-statistic:  5314 on 6 and 12493 DF,  p-value: < 2.2e-16
```

Again there are weak but significant spurious effects of lag climate.

Now consider variation in both a and b.

```
b <- rbeta(n, 1, 2)
a <- rbeta(n,  2, 2) # introduce noise in a
X <- matrix(NA, t, n)
X[1,] <- init_size

for ( i in 2:t ){
    mu <- a + b*X[i-1, ] + beta*E[i-1]
    X[i, ] <- rnorm(n,  mu, sig)
}

dat <-
  data.frame(t = 1:t , size = X, E = E) %>%
  gather( id, size, starts_with('size')) %>%
  mutate( id = as.numeric(factor( id )))

dat <-
  dat %>%
  group_by( id ) %>%
  arrange( id, t ) %>%
  mutate(size0 = lag(size), lag1 = lag( E ), lag2 = lag( E, 2 ), lag3 = lag(E, 3), lag4 = lag(E, 4)) %>%
  mutate( log_growth = size - size0) %>%
  mutate( rll_E = rollmean( lag( E, 2), k = 4, align = 'right', fill = NA))

dat_long <-
  dat %>%
  ungroup() %>%
  filter( complete.cases(.)) %>%
  gather( lag, Env, starts_with('lag'), rll_E)

lag_cors <-
  dat_long %>%
  group_by( lag ) %>%
  summarise( r = round( cor( log_growth, Env ), 2)) %>%
  mutate( label = paste0( 'r = ', r), pos_x = 0.5, pos_y = max(dat$log_growth, na.rm = T )*0.95)

subsamp <- dat_long %>% group_by( t ) %>% sample_frac(0.1, replace = F) # subsample to draw fewer point

dat_long %>%
  ggplot( aes( x = Env, y = log_growth )) +
  geom_point( data = subsamp  , aes( x = Env, y= log_growth ), alpha = 0.1) +
  geom_smooth( method = 'lm', se = F) +
  geom_text( data = lag_cors, aes( label = label, x = pos_x, y = pos_y), hjust = 1) +
  facet_wrap( ~ lag, nrow = 1, scales = 'free_x')
```




```r
m1 <- lm( data = dat, size ~ size0 + lag1 + lag2 + lag3 + lag4 + rll_E )
summary(m1)
```

```
##
## Call:
## lm(formula = size ~ size0 + lag1 + lag2 + lag3 + lag4 + rll_E,
##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31547 -0.24910  0.00016  0.24808  1.44112
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.089814   0.005200  17.273  < 2e-16 ***
## size0        0.897692   0.004200 213.717  < 2e-16 ***
## lag1         0.251185   0.003802  66.059  < 2e-16 ***
## lag2        -0.135147   0.005025 -26.894  < 2e-16 ***
## lag3        -0.031083   0.004946  -6.284  3.4e-10 ***
## lag4        -0.002521   0.004963  -0.508    0.611
## rll_E       -0.019497   0.015092  -1.292    0.196
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.3741 on 12493 degrees of freedom
##   (2500 observations deleted due to missingness)
## Multiple R-squared:  0.8019, Adjusted R-squared:  0.8018
## F-statistic:  8427 on 6 and 12493 DF,  p-value: < 2.2e-16
```

In these simulations there is individual variation in the growth process, this variation is not included in the model however. This leads to weak but often statistically significant negative lag climate climate effects in the fitted model. We could account for individual variation in the growth process by fitting random effects for each plant.

```
mer1 <- lmer( data = dat, size ~ size0 + lag1 + lag2 + lag3 + rll_E +  (size0|id))
summary(mer1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: size ~ size0 + lag1 + lag2 + lag3 + rll_E + (size0 | id)
##    Data: dat
##
## REML criterion at convergence: 7462.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3793 -0.6513  0.0002  0.6422  3.9187
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 0.04024  0.2006
##           size0       0.05571  0.2360   -0.12
##  Residual             0.09044  0.3007
## Number of obs: 12500, groups:  id, 500
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  0.491372   0.011436  42.968
## size0        0.318388   0.013329  23.886
## lag1         0.249748   0.003048  81.938
## lag2         0.001522   0.003795   0.401
## lag3        -0.004679   0.003492  -1.340
## rll_E        0.009645   0.009205   1.048
##
## Correlation of Fixed Effects:
##       (Intr) size0  lag1   lag2   lag3
## size0 -0.397
## lag1  -0.023  0.012
## lag2   0.252 -0.290  0.046
## lag3   0.059 -0.067  0.222  0.322
## rll_E -0.043 -0.042 -0.170 -0.428 -0.493
```

This appears to correct for some of the negative lag effects. But in this case we have many repeated observations of the same plants and each plant has fixed a and b. In reality most plants are not observed more than a few years. Moreover the assumption that a and b are fixed for each plant over time may not be correct.

```r
m1 <- lm( data = dat, size ~ size0 + lag1 + lag2 + lag3 + lag4 + rll_E )
summary(m1)
```

```
##
## Call:
## lm(formula = size ~ size0 + lag1 + lag2 + lag3 + lag4 + rll_E,
##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31547 -0.24910  0.00016  0.24808  1.44112
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.089814   0.005200  17.273  < 2e-16 ***
## size0        0.897692   0.004200 213.717  < 2e-16 ***
## lag1         0.251185   0.003802  66.059  < 2e-16 ***
## lag2        -0.135147   0.005025 -26.894  < 2e-16 ***
## lag3        -0.031083   0.004946  -6.284  3.4e-10 ***
## lag4        -0.002521   0.004963  -0.508    0.611
## rll_E       -0.019497   0.015092  -1.292    0.196
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3741 on 12493 degrees of freedom
##   (2500 observations deleted due to missingness)
## Multiple R-squared:  0.8019, Adjusted R-squared:  0.8018
## F-statistic:  8427 on 6 and 12493 DF,  p-value: < 2.2e-16
```