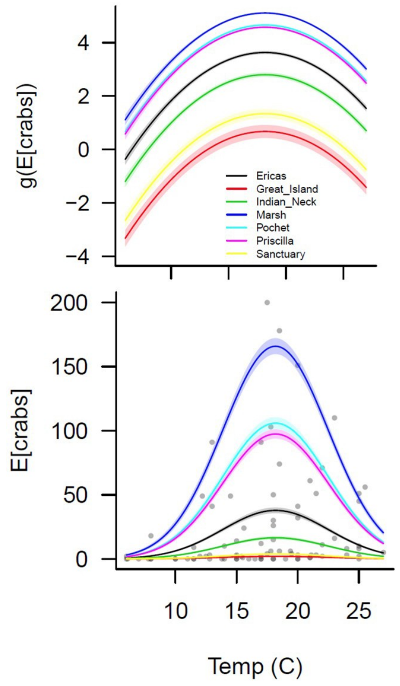
ECO 636
Applied Ecological Statistics

Week 5 – Collinearity



Meg Graham MacLean, PhD

Department of Environmental Conservation

mgmaclean@umass.edu

2021 - Spring

The Week

Monday
- Extra lab day!

Tuesday
- Collinearity

Wednesday (Lab)
- Two-way ANOVA/regression

Thursday
- ANCOVA

Review!

So far…

Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`
Continuous	Single two-level factor	*t*-test	`lm(y~x)`
Continuous	Single multi-level factor	One-way ANOVA	`lm(y~x)`
Continuous	>1 multi-level factor (*)	Two-way ANOVA	lm(y~x_1*x_2)
Continuous	Single continuous	Simple linear regression	`lm(y~x)`

Review!

Next!

Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`
Continuous	Single two-level factor	*t*-test	`lm(y~x)`
Continuous	Single multi-level factor	One-way ANOVA	`lm(y~x)`
Continuous	>1 multi-level factor (*)	Two-way ANOVA	$\texttt{lm(y\sim x}_1\texttt{*x}_2\texttt{)}$
Continuous	Single continuous	Simple linear regression	`lm(y~x)`
Continuous	Multiple continuous	Multiple linear regression	$\texttt{lm(y\sim x}_1\texttt{*x}_2\texttt{)}$

Estimating the relationship with multiple explanatory variables!

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

Multiple linear regression

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

- β_0 is the intercept
- β_1 is the slope of the X_1 relationship
  - The change in $\hat{y}_i$ with one unit change in X_1 at *any* value of X_2 (*additive model!*)
- β_2 is the slope of the X_2 relationship
  - The change in $\hat{y}_i$ with one unit change in X_2 at *any* value of X_1 (*additive model!*)

The changes in $\hat{y}_i$ when with the change in one explanatory variable with the other(s) held constant are often called: marginal effects

Example

Indigo snakes

- 92 indigo snakes
- We are interested in the variation in home range sizes of the snakes (hr.size)
  - We log transformed the home range data (log.HR)
- Our covariates are surrounding habitat structure (proportion)
  - Urban1.50
  - Upland1.50
  - Wetland1.50

Example

5. Evaluate the output

- Model selection
  - What is the best model? What other candidate models are there?

```
> UrUpWe<- lm(log.HR ~ urban1.50  + upland1.50 + wetland1.50, data = indigos)
> UrUp  <- lm(log.HR ~ urban1.50  + upland1.50,  data = indigos)
> UrWe  <- lm(log.HR ~ urban1.50  + wetland1.50, data = indigos)
> UpWe  <- lm(log.HR ~ upland1.50 + wetland1.50, data = indigos)
> Ur    <- lm(log.HR ~ urban1.50,   data=indigos)
> Up    <- lm(log.HR ~ upland1.50,  data=indigos)
> We    <- lm(log.HR ~ wetland1.50, data=indigos)
> m0    <- lm(log.HR ~ 1, data=indigos)
```

How do we pick?

Example

5. Evaluate the output

- Model selection
  - What is the best model? What other candidate models are there?

```
> (modtab <- aictab(fitList))

  Model selection based on AICc:

         K   AICc Delta_AICc AICcWt Cum.Wt      LL
  UrUp   4 236.29       0.00   0.38   0.38 -113.91
  Ur     3 237.15       0.87   0.25   0.63 -115.44
  UrUpWe 5 237.26       0.98   0.24   0.87 -113.28
  UrWe   4 238.48       2.19   0.13   1.00 -115.01
  UpWe   4 246.53      10.25   0.00   1.00 -119.04
  Up     3 250.30      14.01   0.00   1.00 -122.01
  We     3 259.50      23.21   0.00   1.00 -126.61
  m0     2 264.72      28.43   0.00   1.00 -130.29
```

Huh, looks like maybe we should try the Urban and Upland model

Example

Modeling process:

1. *State the question/ hypothesis*
   - *What is the question?*
   - *What are the variables (response and explanatory)?*
2. *Data exploration*
3. *Describe the model*
   - *In word form (should come from your question)*
   - *In mathematical form*
   - *Identify the assumptions of the model*
4. Fit the model! (In R, of course ☺)
5. Evaluate the output
   - Model validation
   - Model selection
6. Interpret the results

Example

4. Fit the model

- Algebra: $y_i = \beta_0 + \beta_1 X_{urbi} + \beta_2 X_{upi} + e_i$
- R:

```
> mTop <- lm(log.HR ~ urban1.50 + upland1.50, data = indigos)
```

Example

4. Fit the model

$$y_i = \beta_0 + \beta_1 X_{urbi} + \beta_2 X_{upi} + e_i$$

```
> summary(mTop)

Call:
lm(formula = log.HR ~ urban1.50 + upland1.50, data = indigos)

Residuals:
     Min       1Q   Median       3Q      Max
-2.59312 -0.64147  0.03546  0.59221  1.83352

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.8548     0.2706  17.941  < 2e-16 ***
urban1.50    -2.0704     0.5001  -4.140 7.89e-05 ***
upland1.50    0.8634     0.4981   1.733   0.0865 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8486 on 89 degrees of freedom
Multiple R-squared:  0.2996,  Adjusted R-squared:  0.2838
F-statistic: 19.03 on 2 and 89 DF,  p-value: 1.316e-07
```

Example

Modeling process:

1. *State the question/ hypothesis*
   - *What is the question?*
   - *What are the variables (response and explanatory)?*
2. *Data exploration*
3. *Describe the model*
   - *In word form (should come from your question)*
   - *In mathematical form*
   - *Identify the assumptions of the model*
4. *Fit the model! (In R, of course ☺)*
5. Evaluate the output
   - Model validation
   - Model selection
6. Interpret the results

Example

5. Evaluate the output

- Model validation - check assumptions!
  - Residuals are normally distributed
  - Constant variance (homogeneity)
  - Observations are independent
  - Predictors measured without error (fixed X)

```
> mTop <- lm(log.HR ~ urban1.50 + upland1.50, data = indigos)
> plot(mTop)
```

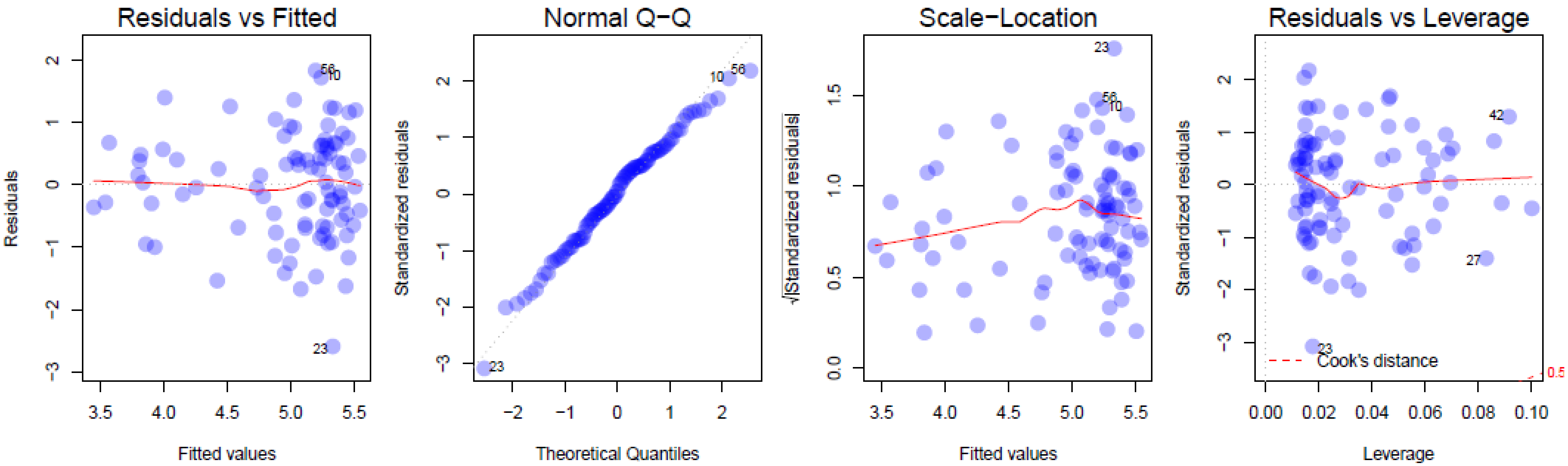
Example

5. Evaluate the output

- Model validation - check assumptions!

```
> mTop <- lm(log.HR ~ urban1.50 + upland1.50, data = indigos)
> plot(mTop)
```

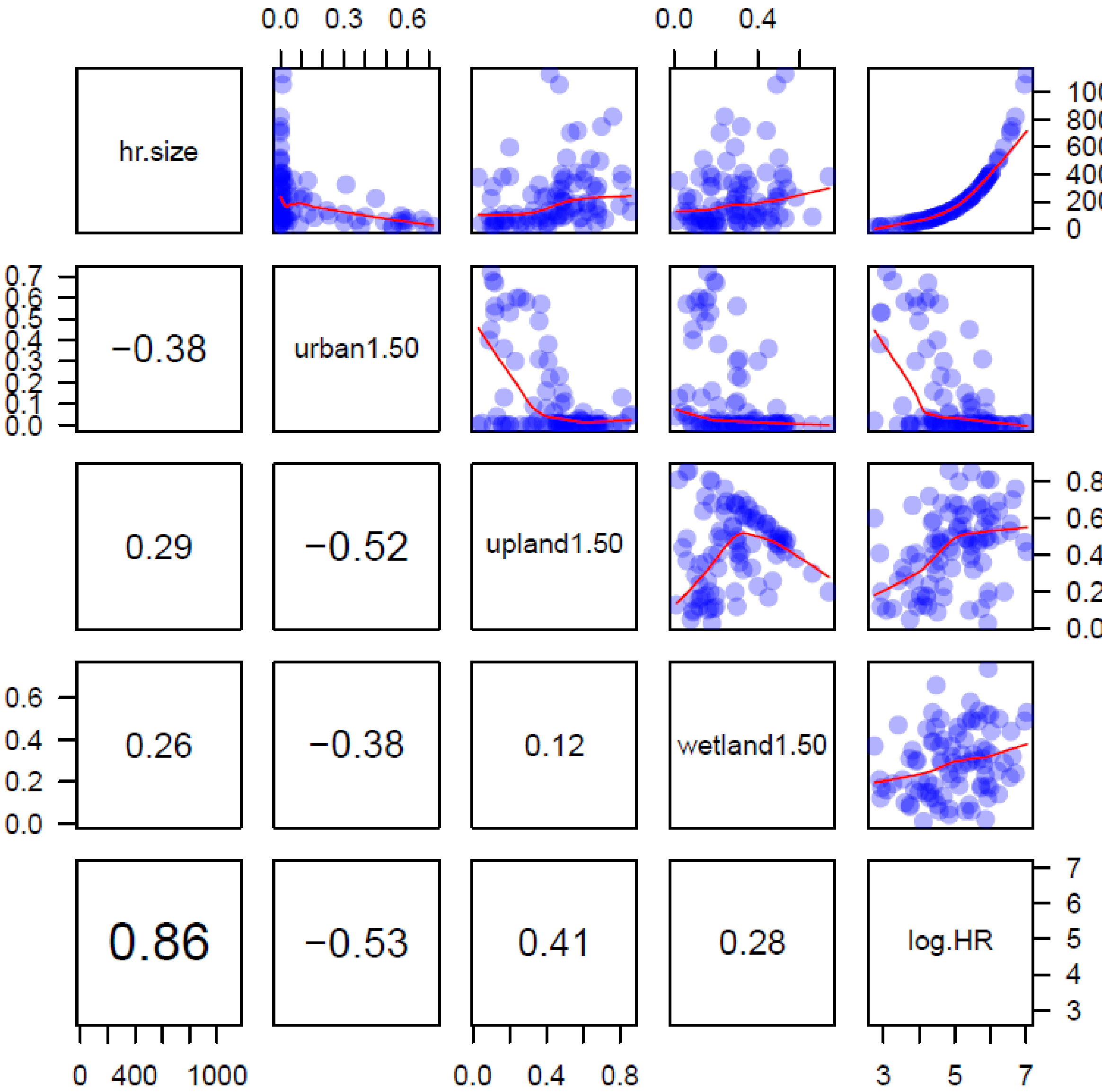
Do we meet our assumptions?



Example

5. Evaluate the output

- Model validation - check assumptions!
  - Residuals are normally distributed
  - Constant variance (homogeneity)
  - **Observations are independent**
  - Predictors measured without error



Some evidence of *collinearity* - we'll deal with this a bit later!

Example

5. Evaluate the output

- Model selection
  - Is this the best model?

What do you think?

```
> summary(mTop)$coefficients
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  4.854762  0.2705938 17.941140 2.600465e-31
urban1.50   -2.070404  0.5001409 -4.139642 7.885350e-05
upland1.50   0.863374  0.4981017  1.733329 8.650048e-02
```

X_{1i}	β_1	p-value
Urban 1.5km	-2.53	<0.0001
Upland 1.5km	1.94	<0.0001
Wetland 1.5km	1.76	0.007

Example

Modeling process:

1. *State the question/ hypothesis*
   - *What is the question?*
   - *What are the variables (response and explanatory)?*
2. *Data exploration*
3. *Describe the model*
   - *In word form (should come from your question)*
   - *In mathematical form*
   - *Identify the assumptions of the model*
4. *Fit the model! (In R, of course ☺)*
5. *Evaluate the output*
   - *Model validation*
   - *Model selection*
6. Interpret the results

Example

6. Interpret the results

- How do we visualize our model?

$$y_i = \beta_0 + \beta_1 X_{urbi} + \beta_2 X_{upi} + e_i$$

```
> summary(mTop)$coefficients
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept)   4.854762  0.2705938 17.941140 2.600465e-31
urban1.50    -2.070404  0.5001409 -4.139642 7.885350e-05
upland1.50    0.863374  0.4981017  1.733329 8.650048e-02
```

$$y_i = 4.85 - 2.07 X_{urbi} + 0.86 X_{upi}$$

- We can visualize our model by holding all but one X constant
  - Typically we use the mean of Xs we are holding constant

Example

6. Interpret the results

- Let's visualize the Urban relationship first
  - $\bar{X}_{up} = 0.44$

$$y_i = 4.85 - 2.07X_{urbi} + 0.86 \times 0.44$$

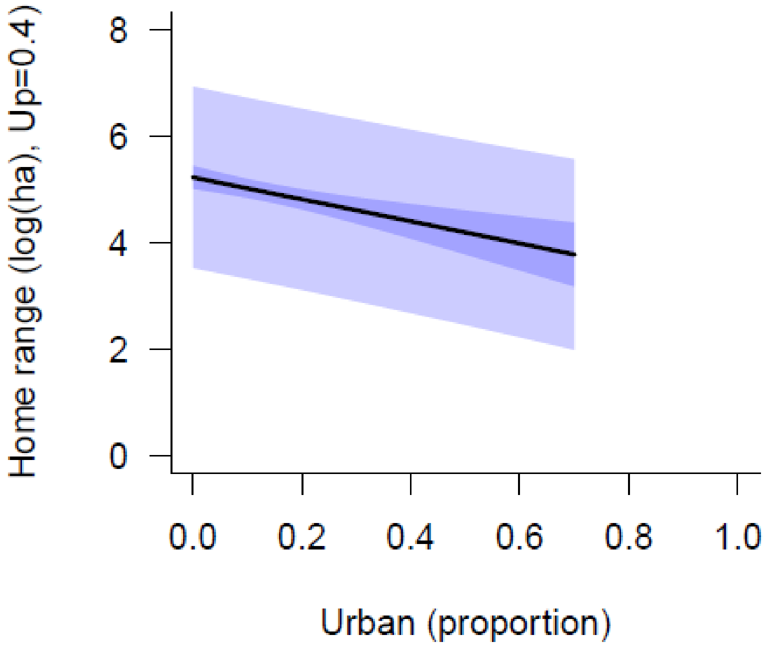
```
> urb.seq <- seq(0,1,0.05) #these are porportions
> up.mean <- mean(indigos$upland1.50)
>
> #vary Urban keep Upland constant
> urb.df <- data.frame(urban1.50 = urb.seq, upland1.50 = up.mean)
```

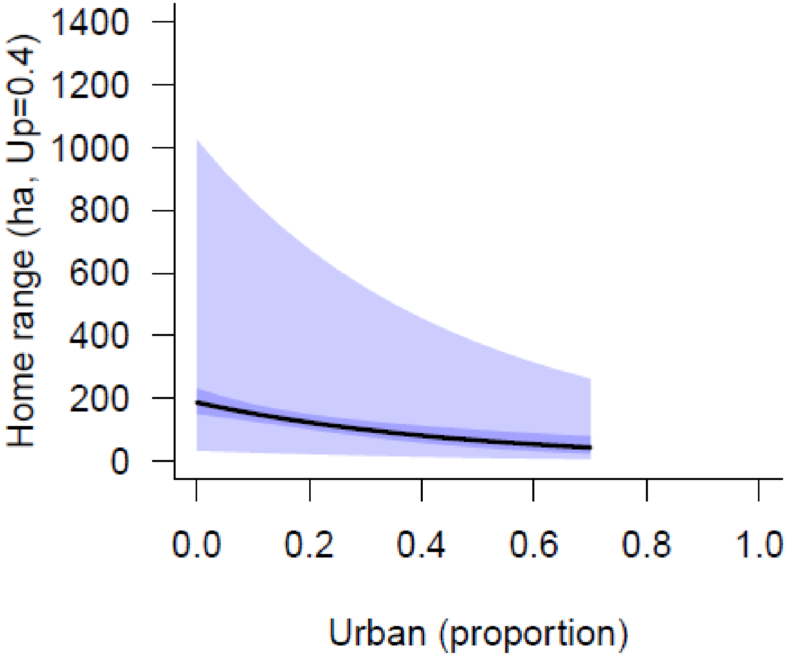
Example

$$y_i = 4.85 - 2.07X_{urbi} + 0.86 \times 0.44$$

- Show relationship and uncertainty in relationship

```
> #Predict Urban relationship
> CI.urb <- predict(mTop, newdata=urb.df, interval="confidence")
> PI.urb <- predict(mTop, newdata=urb.df, interval="prediction")
```





Example

6. Interpret the results

- Let's visualize the Upland relationship next
  - $\bar{X}_{urb} = 0.13$

$$y_i = 4.85 - 2.07 \times 0.13 + 0.86X_{upi}$$

```
> up.seq <- seq(0,1,0.05) #these are porportions
> urb.mean <- mean(indigos$urban1.50)
>
> #vary Upland keep Urban constant
> up.df <- data.frame(urban1.50 = urb.mean, upland1.50 = up.seq)
```

Example

$$y_i = 4.85 - 2.07 \times 0.13 + 0.86X_{upi}$$

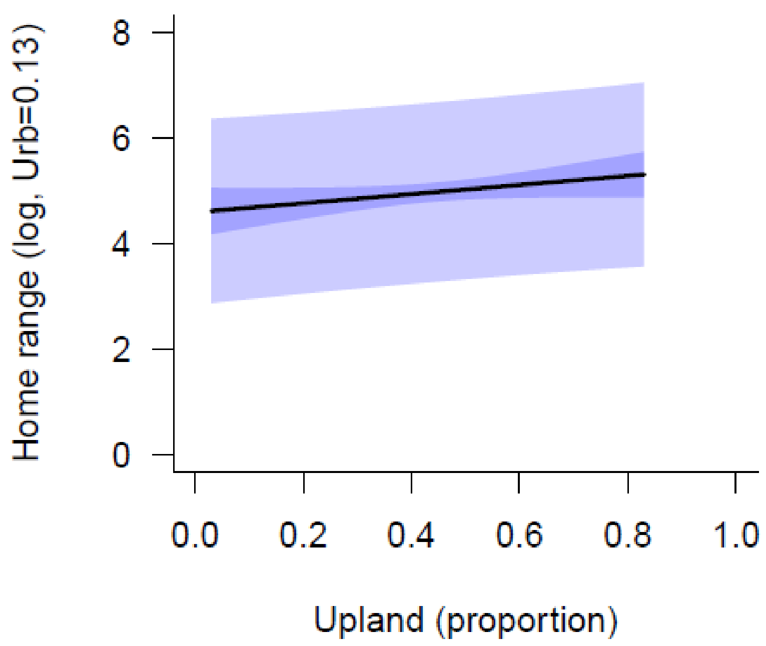
```
> up.seq <- seq(min(indigos$upland1.50),max(indigos$upland1.50),0.05)
> urb.mean <- mean(indigos$urban1.50)
>
> #vary Upland keep Urban constant
> up.df <- data.frame(urban1.50 = urb.mean, upland1.50 = up.seq)
>
> head(up.df)
  urban1.50 upland1.50
1 0.1255435       0.03
2 0.1255435       0.08
3 0.1255435       0.13
4 0.1255435       0.18
5 0.1255435       0.23
6 0.1255435       0.28
```

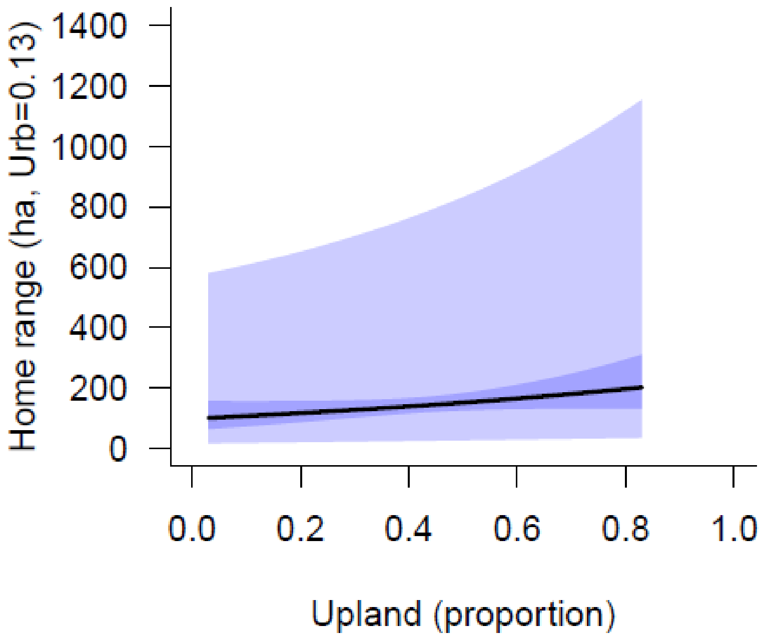
Example

$$y_i = 4.85 - 2.07 \times 0.13 + 0.86X_{upi}$$

- Show relationship and uncertainty in relationship

```
> #Predict Urban relationship
> CI.up <- predict(mTop, newdata=up.df, interval="confidence")
> PI.up <- predict(mTop, newdata=up.df, interval="prediction")
```

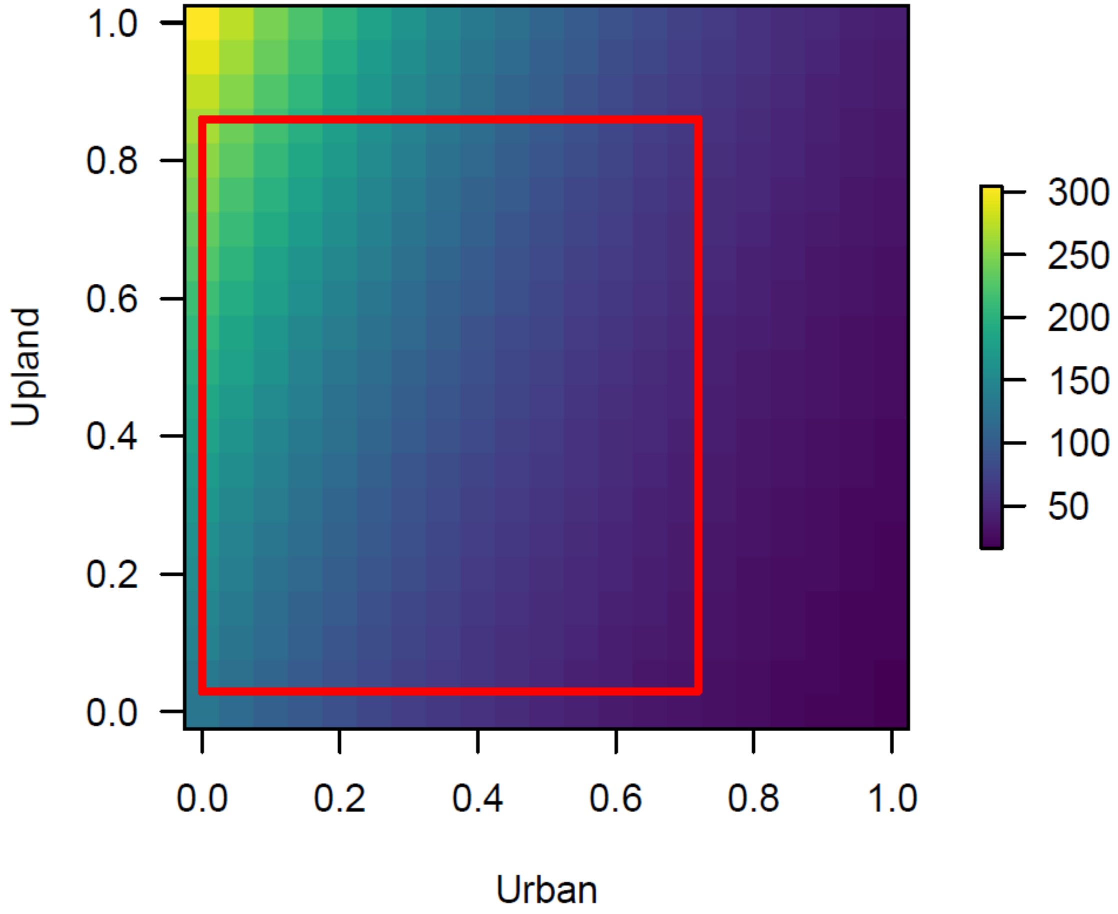




Example

$$y_i = 4.85 - 2.07X_{urbi} + 0.86X_{upi}$$

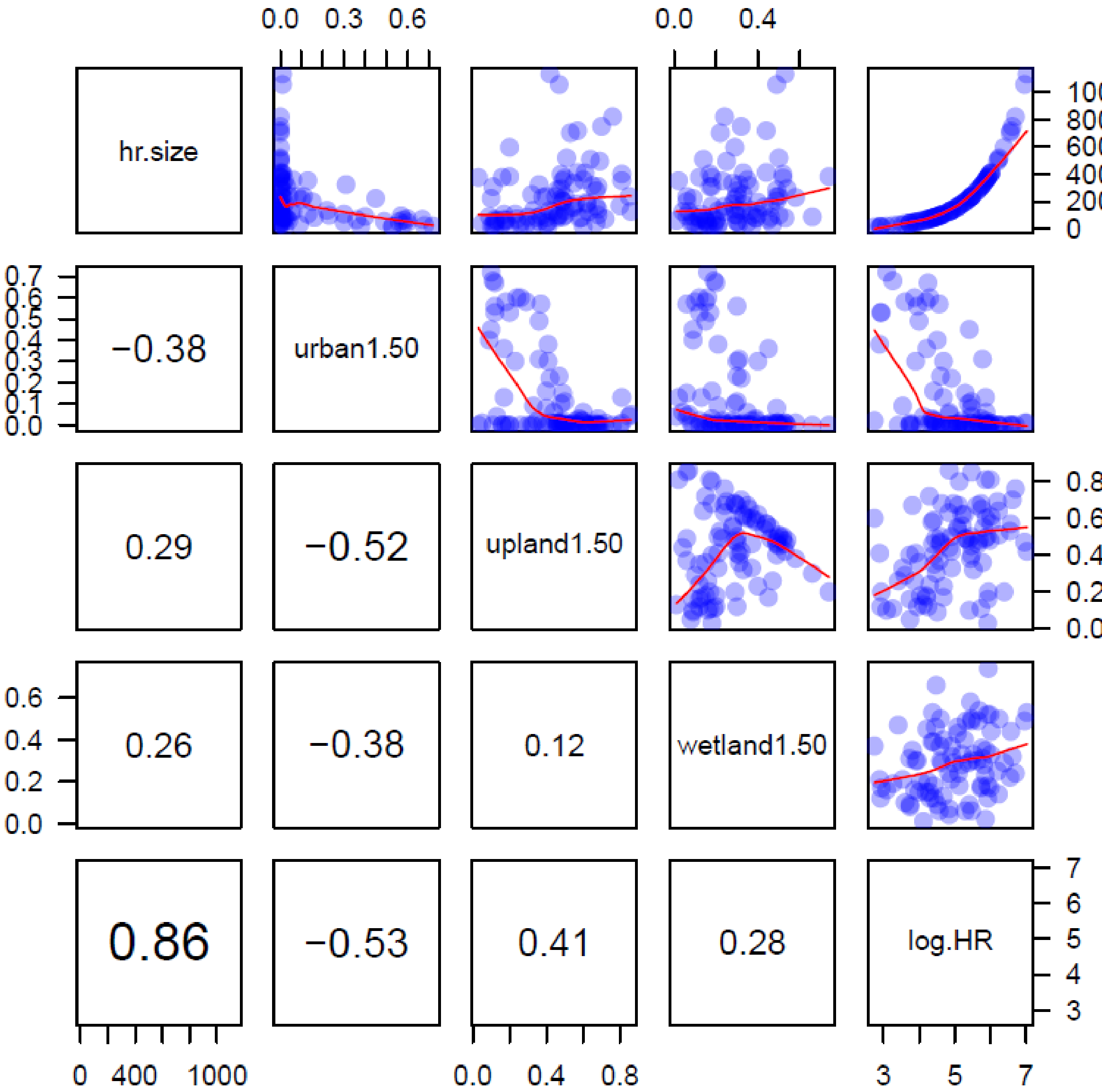
- Can we show the 2-dimensional changes in home range?



Example

5. Evaluate the output

- Model validation - check assumptions!
    - Residuals are normally distributed
    - Constant variance (homogeneity)
    - **Observations are independent**
    - Predictors measured without error

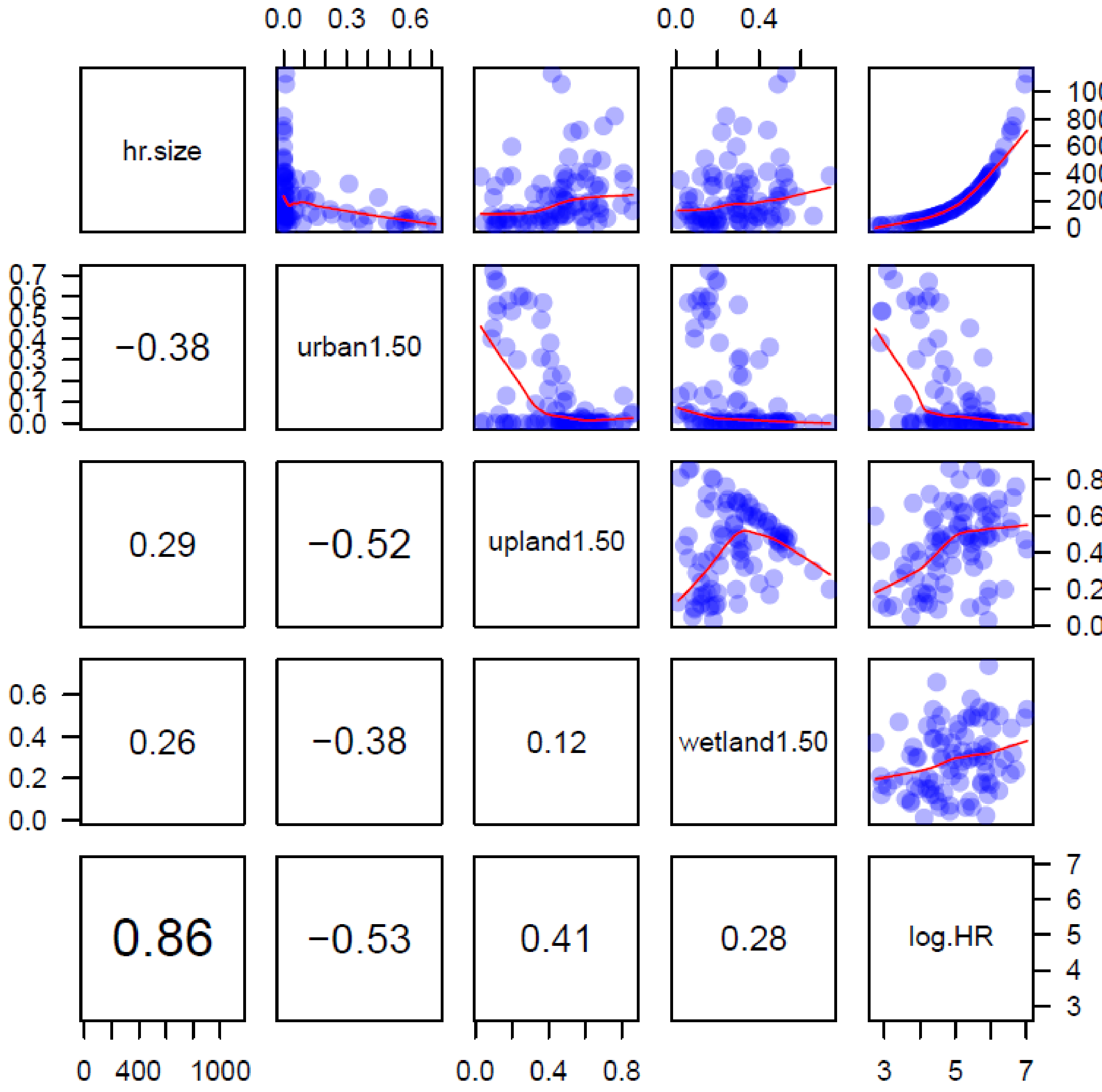


So what about this *collinearity* business? How do we deal with it?

Collinearity

- Correlation between two or more explanatory variables
  - Examples:
    - Sea depth & distance to shore
    - Weight & height
    - Tree diameter & basal area
- What's the issue?
  - Increased Type II errors (false positives)
  - Unstable parameter estimates
    - Variables explaining the same variation

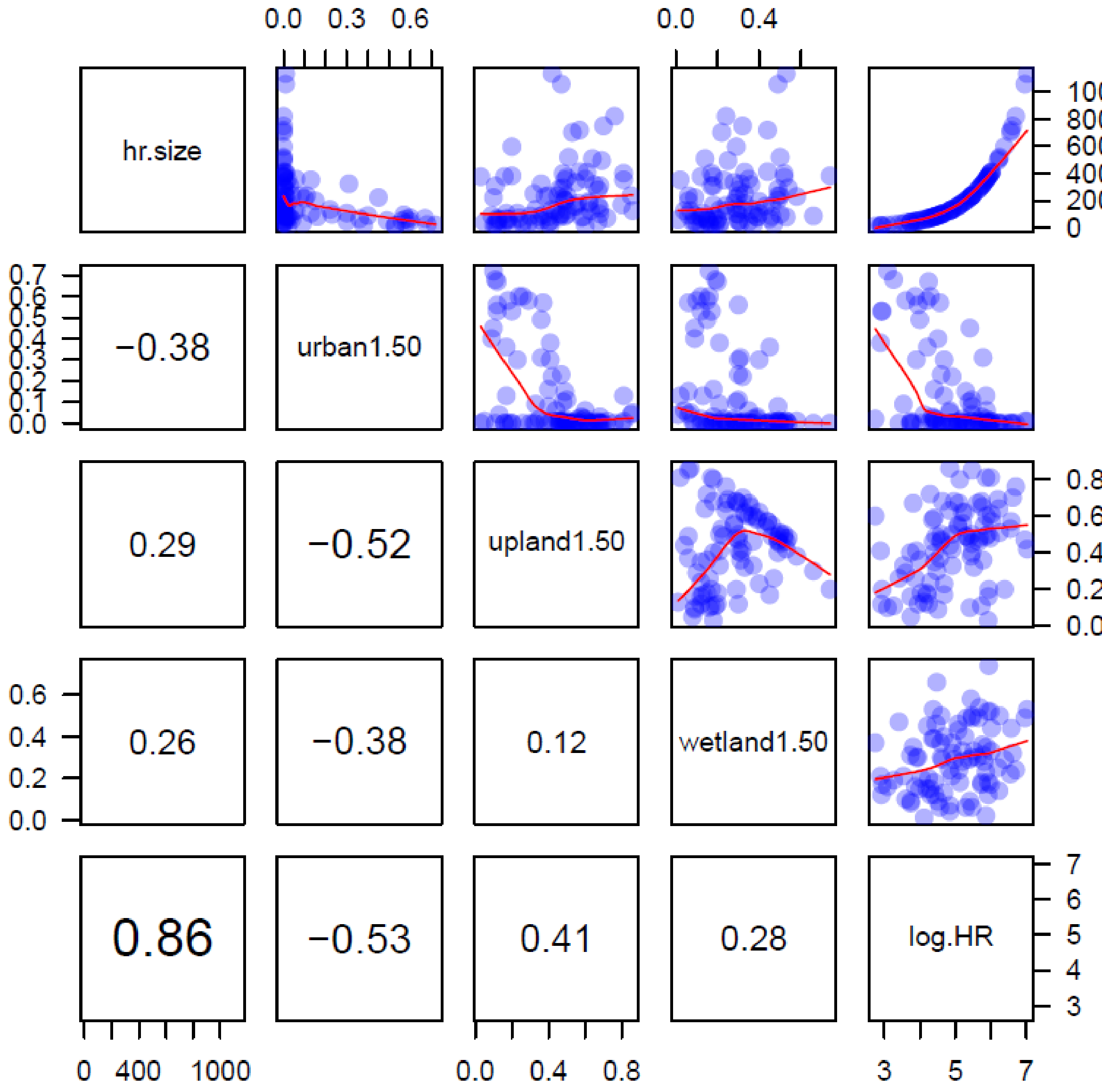
We should identify and remove collinear terms prior to every analysis!



Collinearity

- Can lead to contradictory results!!
  - Dropping a covariate can make non-significant effects significant
  - Sign of estimates can change!

- Collinearity is pretty typical in multiple linear regression – especially as you add explanatory variables.



Example

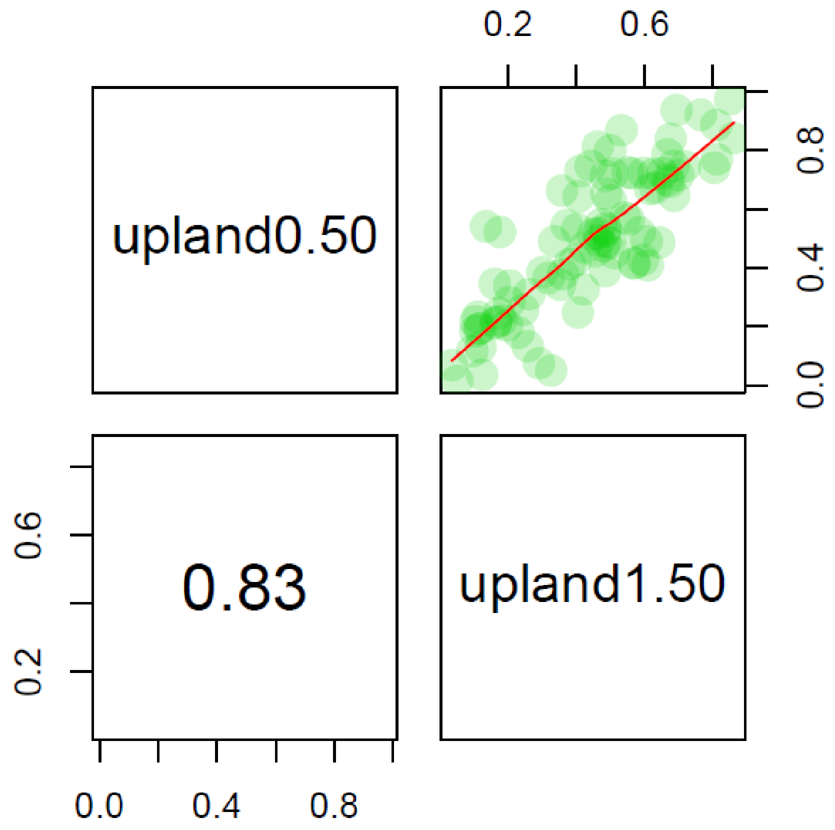
Indigo snakes

- 92 indigo snakes
- We are interested in the variation in home range sizes of the snakes (hr.size)
  - We log transformed the home range data (log.HR)
- Our covariates are surrounding habitat structure (proportion)
  - Urban1.50
  - Upland1.50
  - Wetland1.50
- We actually also have:
  - Urban0.50
  - Upland0.50
  - Wetland0.50

Collinearity

- Let's look at an example with the indigo snakes where we have very high collinearity:

```
> pairs(indigos1[,c("upland0.50","upland1.50")])
```



Collinearity

```
> #fit the upland0.50 model
> m050 <- lm(log.HR ~ upland0.50, data = indigo)
> summary(m050)$coefficients
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 4.252282  0.2276318 18.68053 1.009071e-32
upland0.50  1.453675  0.4147668  3.50480 7.143692e-04
```

```
> #fit the upland0.50 model
> m150 <- lm(log.HR ~ upland1.50, data = indigo)
> summary(m150)$coefficients
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 4.125029  0.2234840 18.45783 2.380743e-32
upland1.50  1.938773  0.4621695  4.19494 6.383963e-05
```

```
> #additive model
> mboth <- lm(log.HR ~ upland0.50 + upland1.50, data = indigo)
> summary(mboth)$coefficients
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 4.1134623  0.2322069 17.714646 6.307482e-31
upland0.50  0.1440552  0.7300513  0.197322 8.440255e-01
upland1.50  1.8020313  0.8343509  2.159800 3.347657e-02
```

Which model might be best?

Collinearity

```
> aictab(list(m050,m150,mboth), modnames = c("m050","m150","mboth"))

  Model selection based on AICc:

        K   AICc Delta_AICc AICcWt Cum.Wt      LL
  m150  3 250.43       0.00   0.69   0.69 -122.08
  mboth 4 252.57       2.15   0.24   0.93 -122.06
  m050  3 255.09       4.66   0.07   1.00 -124.41
```

```
> coef.tab
        (Intercept) upland0.50 upland1.50
  mboth       4.113      0.144      1.802
  m050        4.252      1.454         NA
  m151        4.125         NA      1.939
```

Coefficients are different!

Collinearity

- Collinearity inflates standard errors
  - This is particularly problematic for weak effects
    - Common for observational studies
    - You risk missing important effects!

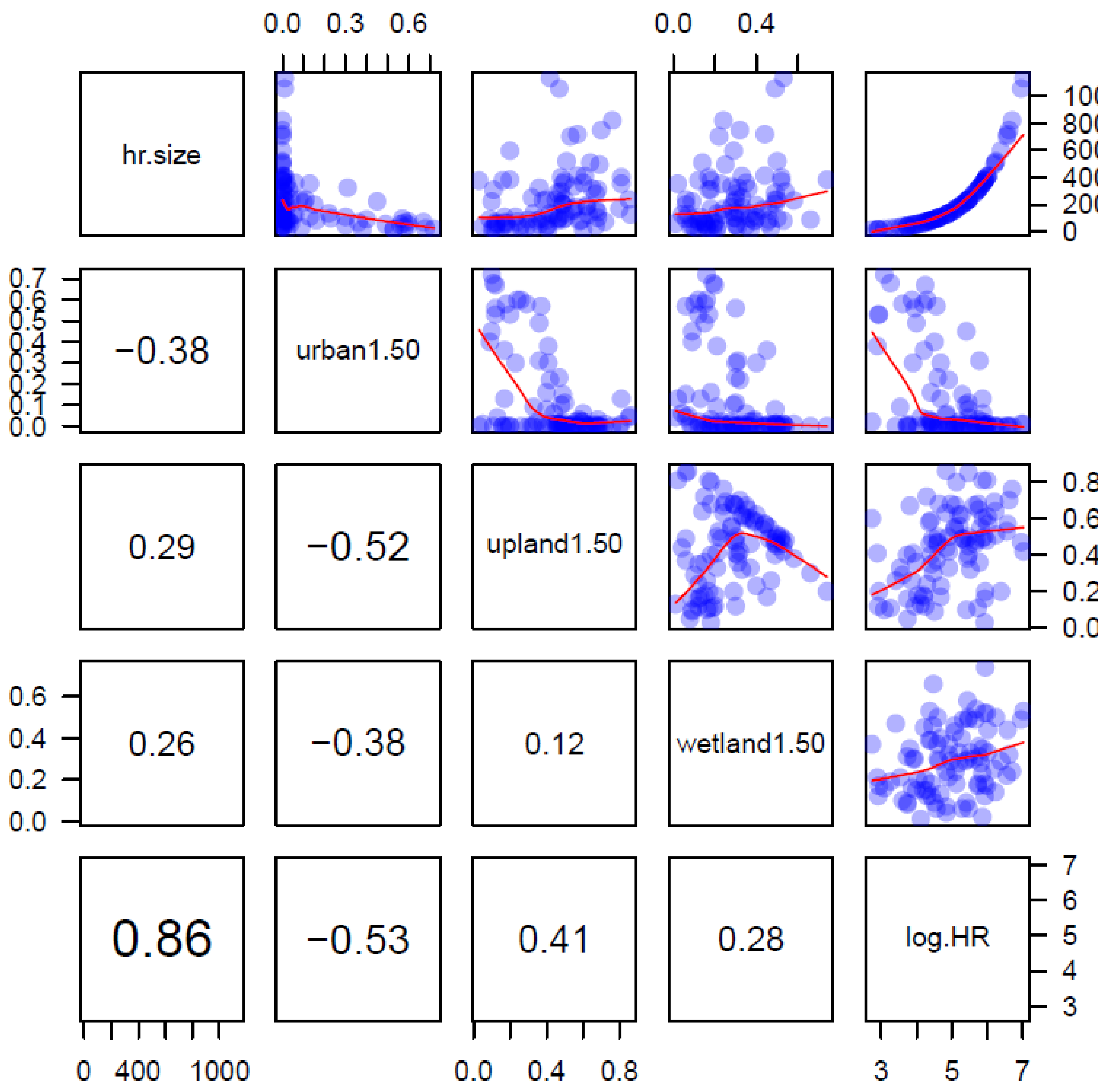
```
> #additive model
> mboth <- lm(log.HR ~ upland0.50 + upland1.50, data = indigo)
> summary(mboth)$coefficients
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 4.1134623  0.2322069 17.714646 6.307482e-31
upland0.50  0.1440552  0.7300513  0.197322 8.440255e-01
upland1.50  1.8020313  0.8343509  2.159800 3.347657e-02
```

So... what do we do about it?

Collinearity

- Remove correlated variables!
  - Rule of thumb: remove variables with $R^2 > 0.6$
  - But… which do you remove?
- You must decide!
  - Keep the most biologically plausible
  - Keep the most interesting relative to your question

WARNING: this only considers pairwise comparisons! We are currently ignoring multidimensional correlation structures!



Collinearity

Variance Inflation Factor (VIF)!

- Quantify the degree of variance inflation
- Removed correlated terms to minimize inflation
- Considers correlation among *all* explanatory variables

$$VIF = \frac{1}{1 - R_j^2}$$

- R_j^2 is the variation in X_j that is explained by all other $X_{i \neq j}$ explanatory variables
- Quantifies the correlation among explanatory variables
  - If explanatory variables are not correlated $R_j^2 = 0$ → no inflation!

How do we calculate R_j^2?

Collinearity

$$VIF = \frac{1}{1 - R_j^2}$$

- R_j^2 is the variation in X_j that is explained by all other $X_{i \neq j}$ explanatory variables
- How to do it:
  - Fit the full model, but with X_j as the *response* - do this for all covariates!
  - Remove terms with *high* VIF values

Collinearity

- In R…

```
> summary(vif.mod <- lm(upland0.50 ~ upland1.50, data = indigo))

  Call:
  lm(formula = upland0.50 ~ upland1.50, data = indigo)

  Residuals:
       Min        1Q    Median        3Q       Max 
  -0.33916 -0.06387 -0.01454  0.07440  0.33335 

  Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
  (Intercept)  0.08029    0.03244   2.475   0.0152 *  
  upland1.50   0.94923    0.06709  14.149   <2e-16 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 0.1339 on 90 degrees of freedom
  Multiple R-squared:  0.6899,	Adjusted R-squared:  0.6864 
  F-statistic: 200.2 on 1 and 90 DF,  p-value: < 2.2e-16
```

Collinearity

- In R… does order matter?

```
> summary(vif.mod <- lm(upland1.50 ~ upland0.50, data = indigo))

  Call:
  lm(formula = upland1.50 ~ upland0.50, data = indigo)

  Residuals:
       Min        1Q    Median        3Q       Max
  -0.33588 -0.07146  0.00706  0.06975  0.23445

  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept)  0.07703    0.02819   2.733  0.00756 **
  upland0.50   0.72675    0.05136  14.149  < 2e-16 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 0.1171 on 90 degrees of freedom
  Multiple R-squared:  0.6899,  Adjusted R-squared:  0.6864
  F-statistic: 200.2 on 1 and 90 DF,  p-value: < 2.2e-16
```

Collinearity

- In R…

$$VIF = \frac{1}{1 - R_j^2}$$

```
> (up.vif <- 1 / (1 - summary(vif.mod)$r.squared))
  [1] 3.224277
```

There must be an easier way…

Collinearity

- In R…

$$VIF = \frac{1}{1 - R_j^2}$$

  - Use the `vif()` function from `library(car)`

```
> #additive model
> mboth <- lm(log.HR ~ upland0.50 + upland1.50, data = indigo)
```

```
> #calculate vifs
> vif(mboth)
upland0.50 upland1.50
  3.224277   3.224277
```

What about a more complex model?

Collinearity

- In R…

$$VIF = \frac{1}{1 - R_j^2}$$

```
> mall <- lm(log.HR ~ upland0.50 + urban0.50 + wetland0.50 +
+                     upland1.50 + urban1.50 + wetland1.50, data=indigo)
```

```
> #calculate vifs
> vif(mall)
 upland0.50   urban0.50 wetland0.50  upland1.50   urban1.50 wetland1.50
   5.969110    6.777936    4.131077    4.503149    6.902179    3.009913
```

Let's remove the worst one and try again…

Collinearity

- In R…

$$VIF = \frac{1}{1 - R_j^2}$$

```
> #fit the reduced model
> msub <- lm(log.HR ~ upland0.50 + urban0.50 + wetland0.50 +
+                     upland1.50 +              wetland1.50, data=indigo)
```

```
> #calculate vifs
> vif(msub)
 upland0.50   urban0.50 wetland0.50  upland1.50 wetland1.50 
   4.850618    1.640862    2.872044    3.544600    2.283783 
```

What is too high??

Collinearity

Variance Inflation Factor (VIF)!

- Recommended cutoffs range from 2-5
  - 2 is a more conservative cutoff (allow less collinearity)
  - 5 is a less conservative cutoff (more collinearity allowed)

```
> #fit the reduced model
> msub <- lm(log.HR ~ upland0.50 + urban0.50 + wetland0.50 +
+                     upland1.50 +             wetland1.50, data=indigo)
```

```
> #calculate vifs
> vif(msub)
 upland0.50   urban0.50 wetland0.50  upland1.50 wetland1.50
   4.850618    1.640862    2.872044    3.544600    2.283783
```

Let's try again!

Collinearity

Variance Inflation Factor (VIF)!

- Recommended cutoffs range from 2-5
  - 2 is a more conservative cutoff (allow less collinearity)
  - 5 is a less conservative cutoff (more collinearity allowed)

```
> #fit the reduced model
> msub <- lm(log.HR ~                urban0.50 + wetland0.50 +
+                     upland1.50                    , data=indigo)
```

```
> #calculate vifs
> vif(msub)
  urban0.50 wetland0.50  upland1.50
   1.364749    1.119469    1.337555
```

Better!

Collinearity

Variance Inflation Factor (VIF)!

- Recommended cutoffs range from 2-5
  - 2 is a more conservative cutoff (allow less collinearity)
  - 5 is a less conservative cutoff (more collinearity allowed)
- Variables of interest might have high VIF
  - It is your (the analyst's) choice of which variables to use
  - The main point is to remove collinearity
- If all of your predictors are of interest:
  - Only consider models in which all VIFs < chosen cutoff
  - Don't have multi-scale variables in the same model (e.g., Urban0.50 and Urban1.50)

For next week:

1) Read Fox Ch. 6 for an intro to GLMs
2) Watch the recorded lecture and do the exercise
3) Finish the two-part lab exercise for this week on regression and review of two-way ANOVA
4) Complete the individual assessment on Moodle by 11:55pm Monday night.