

Analisis de los datos de migraciones de la mosca de las frutas

Anna Sikov

UNI

June 8, 2020

El archivo medfly.txt contiene datos coletados para investigar las migraciones de la mosca mediterránea de la fruta (la descripción completa de los datos está en el archivo Medfly descripcion.pdf), donde la pregunta de investigación es si la mosca mediterránea de la fruta pasa el invierno en regiones más frías de Israel o migran a lugares más calientes y regresan cuando pasa el invierno.

Ajustar un modelo lineal para predecir el número de moscas atrapadas (A), dependiendo de la localización de la trampa.

1. Definir el modelo lineal para responder a la pregunta de la investigación.
2. Estimar los coeficientes de la regresión.
3. Calcular intervalos de confianza de 95% para los coeficientes de la regresión.
4. Calcular el R^2 del modelo.

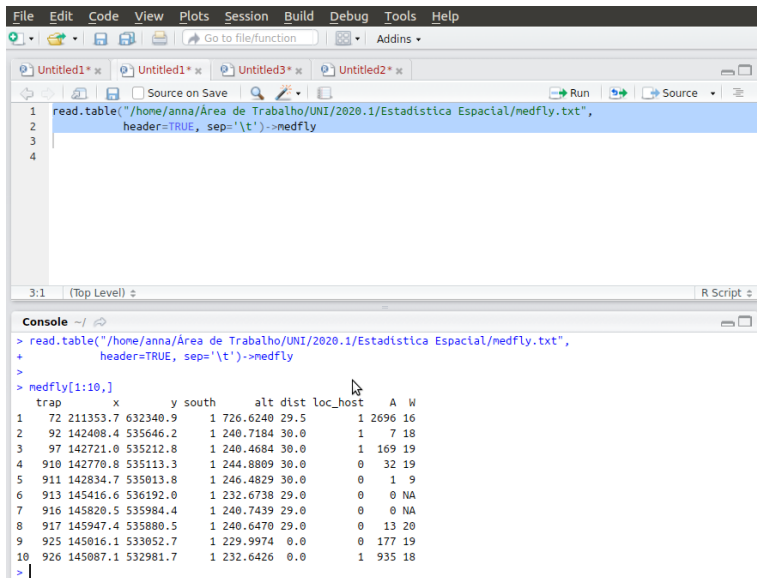
Para este ejercicio ustedes tienen que hacer el cálculo en el software R, sin utilizar las funciones como lm, glm, etc. Para esto ustedes tienen que definir todas las matrices y vectores relevantes y utilizar las fórmulas que ustedes estudiaron en el curso de modelos lineales.

Cómo responderían a la pregunta de la investigación?

2. Descripción de los datos: Fueron analizadas 89 trampas puestas en locales diferentes en Israel (Centro o Sur) por un período de 27 semanas. Para cada trampa tenemos la siguiente información:

- **trap**: El número de identificación de la trampa
- **x**: La coordenada x de la trampa.
- **y**: La coordenada y de la trampa.
- **south**: La variable indicadora (1 si la trampa se localiza en la parte de Sur, y 0 caso contrario).
- **alt**: La altura de la trampa (sobre el nivel del mar).
- **dist**: La distancia de la trampa hasta la región más caliente.
- **loc host**: La variable indicadora 1- si se encuentro un huésped (organismo que alberga a otro en su interior o que lo porta sobre sí) dentro de 50 metros de la trampa, 0- caso contrario.
- **A**: el número de moscas atrapados
- **W**: semana de la primera captura (sólo si $A > 0$).

Paso 1: Abrir el archivo



The screenshot shows the RStudio environment. The script editor contains the following code:

```
1 read.table("/home/anna/Área de Trabalho/UNI/2020.1/Estadística Espacial/medfly.txt",
2           header=TRUE, sep='\t')->medfly
3
4
```

The console shows the execution of the code and the resulting data frame:

```
> read.table("/home/anna/Área de Trabalho/UNI/2020.1/Estadística Espacial/medfly.txt",
+           header=TRUE, sep='\t')->medfly
>
> medfly[1:10,]
  trap      x      y south    alt dist loc_host    A    W
1    72 211353.7 632340.9    1 726.6240 29.5    1 2696 16
2    92 142408.4 535646.2    1 240.7184 30.0    1    7 18
3    97 142721.0 535212.8    1 240.4684 30.0    1 169 19
4   910 142770.8 535113.3    1 244.8809 30.0    0  32 19
5   911 142834.7 535013.8    1 246.4829 30.0    0    1  9
6   913 145416.6 536192.0    1 232.6738 29.0    0    0 NA
7   916 145820.5 535984.4    1 240.7439 29.0    0    0 NA
8   917 145947.4 535880.5    1 240.6470 29.0    0   13 20
9   925 145016.1 533052.7    1 229.9974  0.0    0  177 19
10  926 145087.1 532981.7    1 232.6426  0.0    1  935 18
```

```
1 read.table("/home/anna/Área de Trabalho/UNI/2020.1/Estatística Espacial/medfly.txt",
2           header=TRUE, sep='\t')->medfly
3

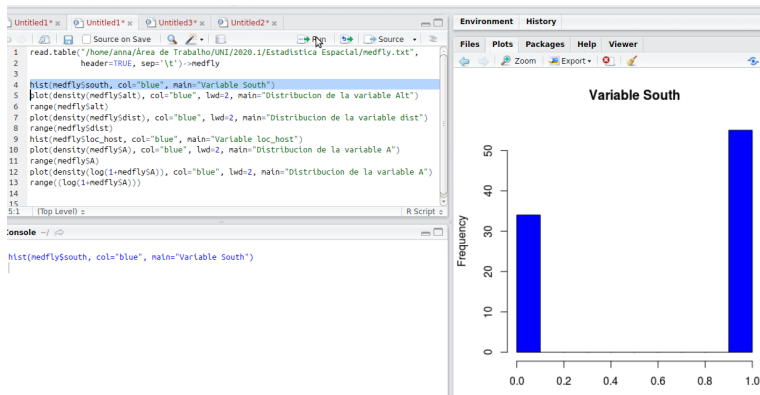
3:1 (Top Level) R Script

Console ~/
>
> medfly[1:10,]
  trap      x      y south      alt dist loc_host      A      W
1    72 211353.7 632340.9      1 726.6240 29.5      1 2696 16
2    92 142408.4 535646.2      1 240.7184 30.0      1    7 18
3    97 142721.0 535212.8      1 240.4684 30.0      1 169 19
4   910 142770.8 535113.3      1 244.8809 30.0      0  32 19
5   911 142834.7 535013.8      1 246.4829 30.0      0    1  9
6   913 145416.6 536192.0      1 232.6738 29.0      0    0 NA
7   916 145820.5 535984.4      1 240.7439 29.0      0    0 NA
8   917 145947.4 535880.5      1 240.6470 29.0      0  13 20
9   925 145016.1 533052.7      1 229.9974  0.0      0 177 19
10  926 145087.1 532981.7      1 232.6426  0.0      1 935 18
>
> medfly$trap
 [1] 72 92 97 910 911 913 916 917 925 926 932 933 935 938 939 1035 1200 1205 1220 1225
[21] 1240 1245 1260 1270 1271 1274 1280 1290 1291 1292 1295 1305 1308 1310 1311 1315 1319 1320 1321 1322
[41] 1323 1325 1326 1327 1328 1334 1337 1342 1344 1362 1383 1384 9103 9105 9106 9112 9115 9118 9119 9123
[61] 9128 9129 9130 9131 9136 9139 9143 9144 9202 9203 9204 9205 9206 9210 9212 9219 9221 9222 9225 9226
[81] 9232 9235 9252 9254 9256 9257 9259 9260 9262
>
> medfly$A
 [1] 2696  7 169 32  1  0  0 13 177 935 33  7  8 769  0 563 1081 611 2684  70
[21] 495 848  54 301 377  5 497 273  36 17 82 812 368 996 251 20 203  6 23 35
[41] 1655 1051 142 190 4 1849 2686 3055 276 616 756 2087 15 23 13 1228 97 205 86 1191
[61] 29  2  8 120 45 37 1018 852  3 10 492 117 908 50 34 42  0 994 155 80
[81]  9  1  8  0 25 221 176 60 343
```

Paso 2: Analisis Preliminar

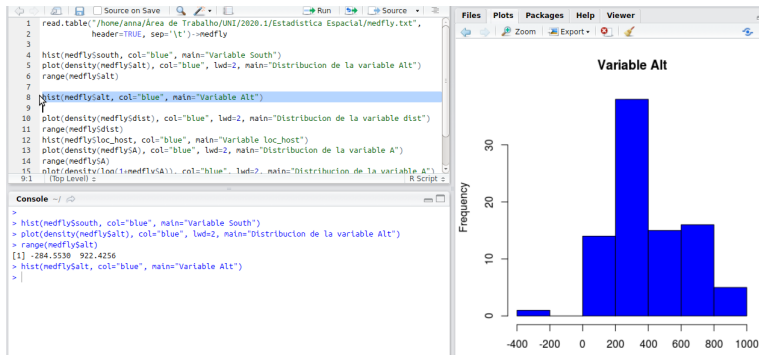
Es necesario hacer analisis preliminar para encontrar errores en los datos, y además para saber qué tipo de modelo vamos a ajustar a los datos

Variable South



Paso 2: Analisis Preliminar

Variable Alt



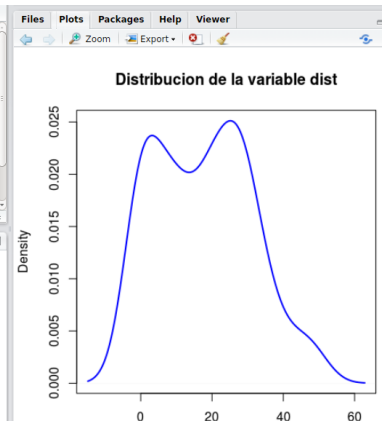
Observen, que hay una observación con un valor negativo. Si es un error en los datos, esta observación no puede ser utilizada para ajustar un modelo! En Israel existen lugares que estan debajo del nivel del mar, entonces no es error.

Paso 2: Analisis Preliminar

Variable dist

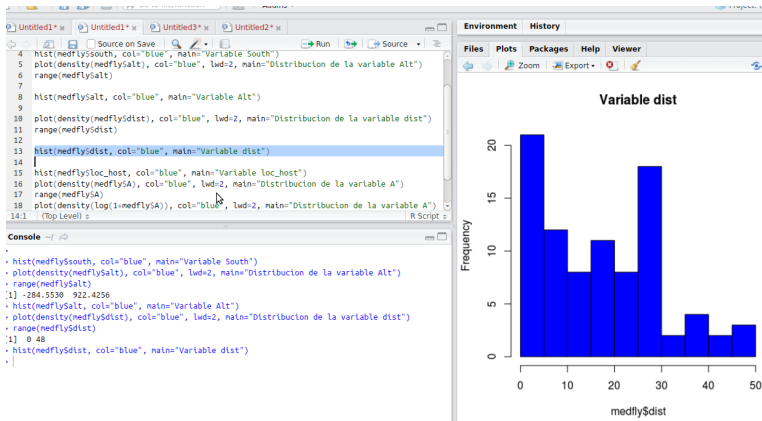
```
1 read.table("/home/anna/Área de Trabalho/UNI/2020.1/Estatística Espacial/medfly.txt",
2             header=TRUE, sep='\t')->medfly
3
4 hist(medfly$South, col="blue", main="Variable South")
5 plot(density(medfly$Salt), col="blue", lwd=2, main="Distribucion de la variable Alt")
6 range(medfly$Salt)
7
8 hist(medfly$Salt, col="blue", main="Variable Alt")
9
10 plot(density(medfly$dist), col="blue", lwd=2, main="Distribucion de la variable dist")
11 range(medfly$dist)
12 hist(medfly$loc_host, col="blue", main="Variable loc_host")
13 plot(density(medfly$A), col="blue", lwd=2, main="Distribucion de la variable A")
14 range(medfly$A)
15 plot(density(log(1+medfly$A)), col="blue", lwd=2, main="Distribucion de la variable A")
12:1 | (Top Level) | R Script
```

```
Console ~/
>
> hist(medfly$South, col="blue", main="Variable South")
> plot(density(medfly$Salt), col="blue", lwd=2, main="Distribucion de la variable Alt")
> range(medfly$Salt)
[1] -284.5530 922.4256
> hist(medfly$Salt, col="blue", main="Variable Alt")
> plot(density(medfly$dist), col="blue", lwd=2, main="Distribucion de la variable dist")
> range(medfly$dist)
[1] 0 48
>
```



Paso 2: Analisis Preliminar

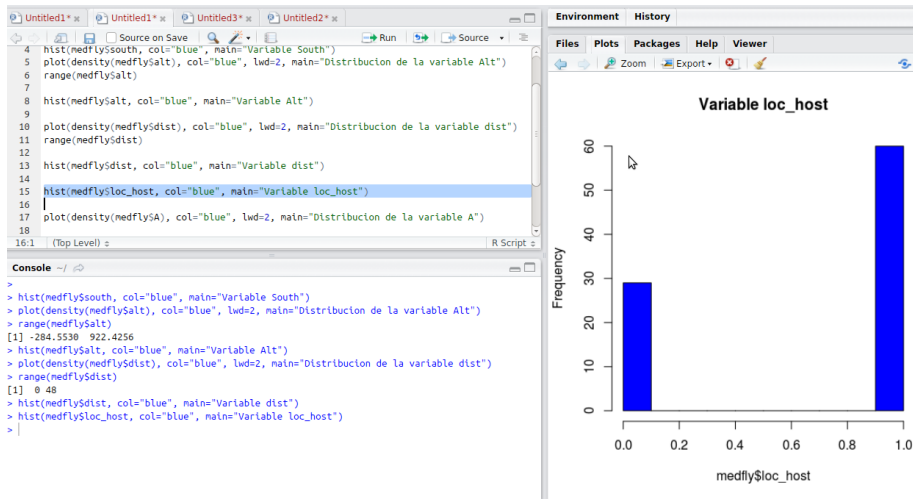
Variable dist



Hay pocas trampas que están a distancia de más de 30 km de las regiones calientes. Esta información puede ser útil cuando ajustamos el modelo o analizamos los resultados.

Paso 2: Analisis Preliminar

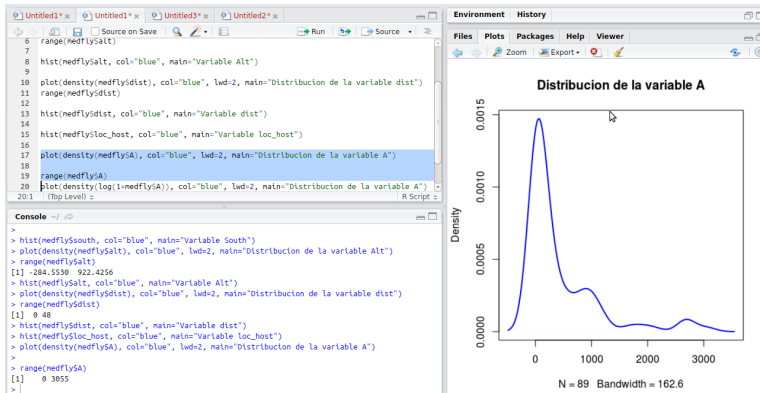
Variable loc host



Esta variable binaria también puede ser utilizada cuando ajustamos un modelo puesto que hay suficientes observaciones con en cada categoría.

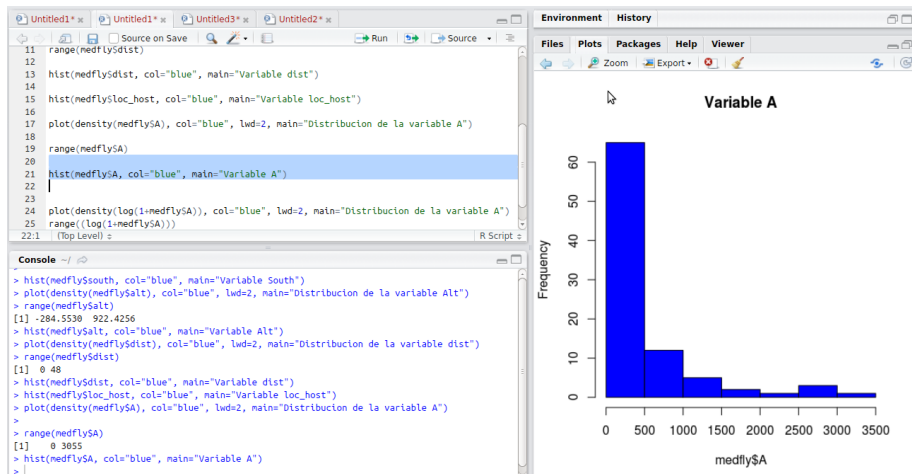
Paso 2: Analisis Preliminar

Variable A- la variable dependiente



Paso 2: Analisis Preliminar

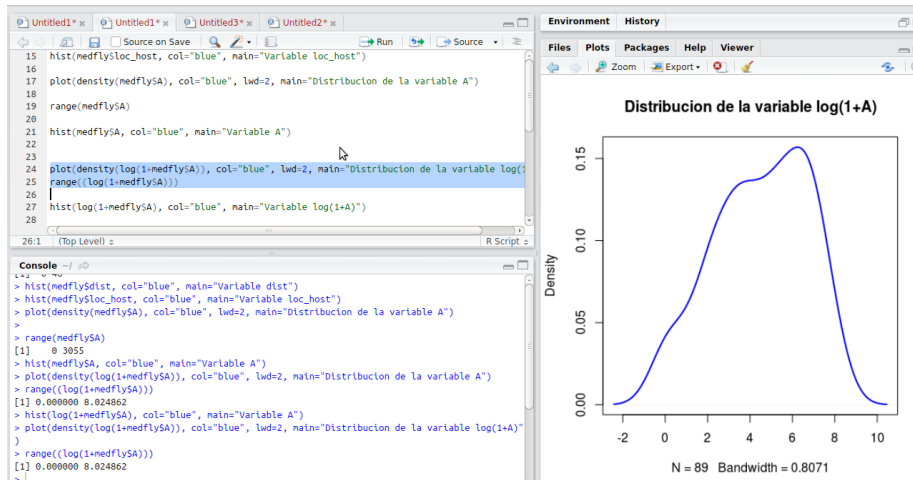
Variable A- la variable dependiente



Parece que aquí tenemos que hacer una transformación (porqué?)

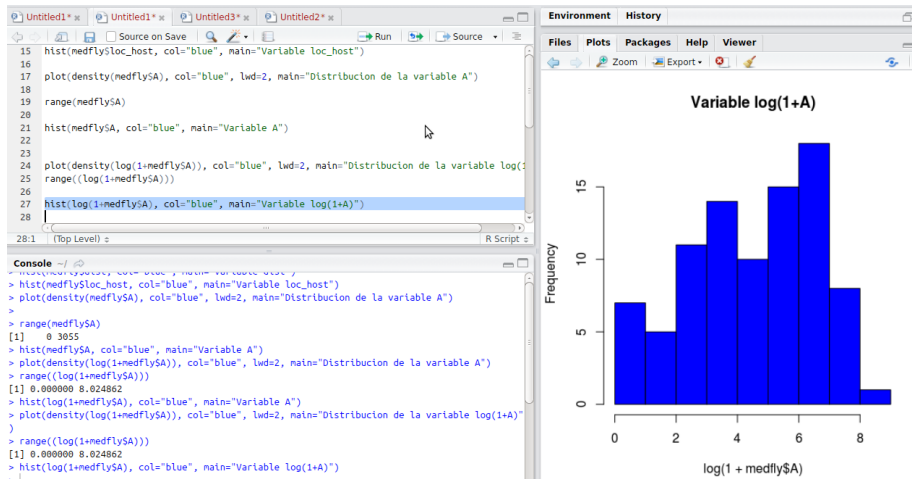
Paso 2: Analisis Preliminar

Variable $\log(A)$ - la variable dependiente



Paso 2: Analisis Preliminar

Variable $\log(1+A)$ - la variable dependiente



Entonces vamos a utilizar la variable $\log(1+A)$ como la variable dependiente. (Porqué no $\log(A)$?)

Paso 2: Analisis Preliminar

```
15 hist(medfly$loc_host, col="blue", main="Variable loc_host")
16
17 plot(density(medfly$A), col="blue", lwd=2, main="Distribucion de la variable A")
18
19 range(medfly$A)
20
21 hist(medfly$A, col="blue", main="Variable A")
22
23
24 plot(density(log(1+medfly$A)), col="blue", lwd=2, main="Distribucion de la variable log(1+A)")
25 range((log(1+medfly$A)))
26
27 hist(log(1+medfly$A), col="blue", main="Variable log(1+A)")
28
```

28:1 (Top Level) ⌵ R Script

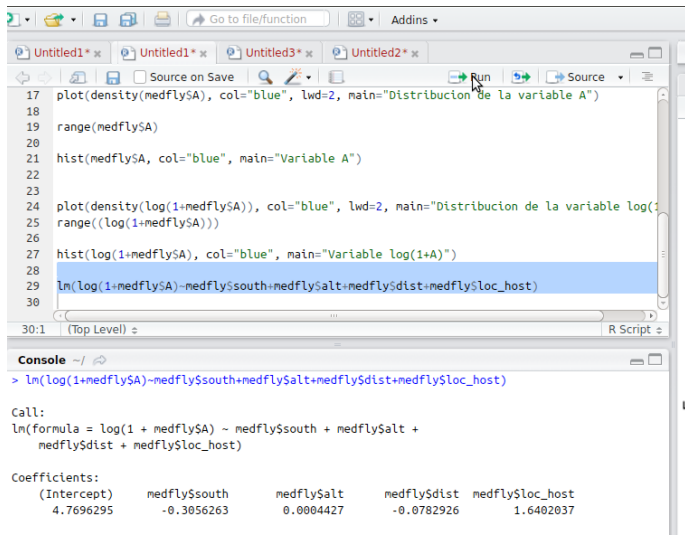
Console ~/ ↶ ⌵ ⌵

```
> plot(density(log(1+medfly$A)), col="blue", lwd=2, main="Distribucion de la variable log(1+A)")
> range((log(1+medfly$A)))
[1] 0.000000 8.024862
> hist(log(1+medfly$A), col="blue", main="Variable A")
> plot(density(log(1+medfly$A)), col="blue", lwd=2, main="Distribucion de la variable log(1+A)")
> range((log(1+medfly$A)))
[1] 0.000000 8.024862
> hist(log(1+medfly$A), col="blue", main="Variable log(1+A)")
>
> medfly$W
[1] 16 18 19 19 9 NA NA 20 19 18 18 17 19 19 NA 8 1 2 9 16 12 9 11 15 14 25 16 18 18
[30] 21 17 17 11 2 0 19 18 15 17 20 15 14 6 10 10 15 13 6 2 17 6 10 17 17 18 6 1 18
[59] 19 17 1 3 21 19 17 18 1 0 26 21 17 18 18 25 19 24 NA 18 19 19 10 21 20 NA 19 18 18
[88] 26 19
>
```

Paso 3: Modelo de Regresión Lineal

$$Y_i = X_i\beta + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$$

Modelo lineal: función "lm" de R



```
17 plot(density(medflySA), col="blue", lwd=2, main="Distribucion de la variable A")
18
19 range(medflySA)
20
21 hist(medflySA, col="blue", main="Variable A")
22
23
24 plot(density(log(1+medflySA)), col="blue", lwd=2, main="Distribucion de la variable log(1+A)")
25 range((log(1+medflySA)))
26
27 hist(log(1+medflySA), col="blue", main="Variable log(1+A)")
28
29 lm(log(1+medflySA)~medfly$south+medfly$salt+medfly$dist+medfly$loc_host)
30
```

Console ~/ ↗

```
> lm(log(1+medflySA)~medfly$south+medfly$salt+medfly$dist+medfly$loc_host)
```

Call:

```
lm(formula = log(1 + medflySA) ~ medfly$south + medfly$salt +
    medfly$dist + medfly$loc_host)
```

Coefficients:

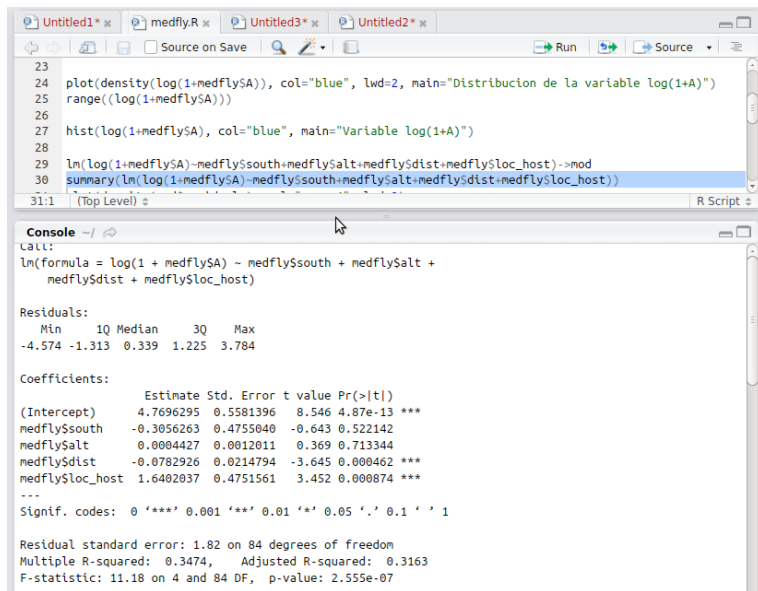
(Intercept)	medfly\$south	medfly\$salt	medfly\$dist	medfly\$loc_host
4.7696295	-0.3056263	0.0004427	-0.0782926	1.6402037

Paso 3: Modelo de Regresión Lineal

$$Y_i = X_i\beta + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$$

Si la pregunta de investigación es si la mosca mediterránea de la fruta pasa el invierno en regiones más frías de Israel o migran a lugares más calientes y regresan cuando pasa el invierno, porque es necesario utilizar todas las variables en el modelo (no solo la variable dist?)?

Paso 3: Modelo de Regresión Lineal



The screenshot shows the R Studio environment. The top pane contains R code for data visualization and model fitting. The bottom pane shows the console output of the fitted model.

```
23
24 plot(density(log(1+medfly$A)), col="blue", lwd=2, main="Distribucion de la variable log(1+A)")
25 range((log(1+medfly$A)))
26
27 hist(log(1+medfly$A), col="blue", main="Variable log(1+A)")
28
29 lm(log(1+medfly$A)~medfly$south+medfly$salt+medfly$dist+medfly$loc_host)->mod
30 summary(lm(log(1+medfly$A)~medfly$south+medfly$salt+medfly$dist+medfly$loc_host))
31:1 (Top Level) ↕
```

Console ~/ ↕

Call:
lm(formula = log(1 + medfly\$A) ~ medfly\$south + medfly\$salt +
medfly\$dist + medfly\$loc_host)

Residuals:

	Min	1Q	Median	3Q	Max
	-4.574	-1.313	0.339	1.225	3.784

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.7696295	0.5581396	8.546	4.87e-13 ***
medfly\$south	-0.3056263	0.4755040	-0.643	0.522142
medfly\$salt	0.0004427	0.0012011	0.369	0.713344
medfly\$dist	-0.0782926	0.0214794	-3.645	0.000462 ***
medfly\$loc_host	1.6402037	0.4751561	3.452	0.000874 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.82 on 84 degrees of freedom
Multiple R-squared: 0.3474, Adjusted R-squared: 0.3163
F-statistic: 11.18 on 4 and 84 DF, p-value: 2.555e-07

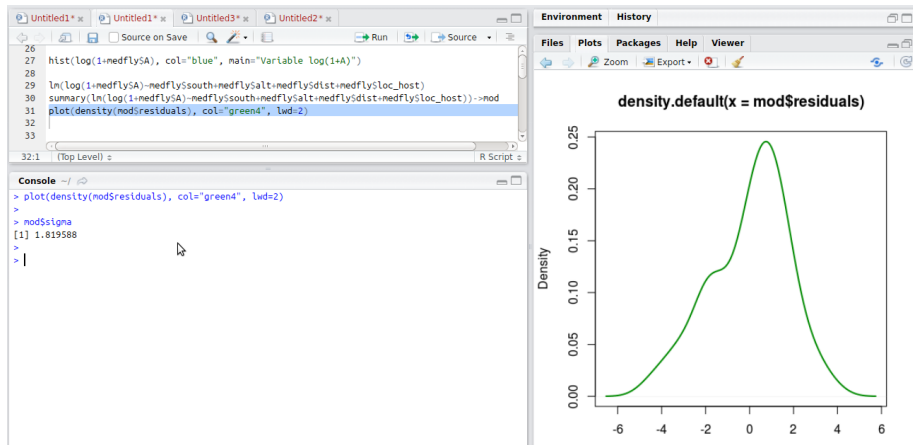
Modelo de Regresión Lineal

$$\hat{Y}_i = 4.76 - 0.306S_i + 0.00044Alt_i - 0.0782Dist_i + 1.64H_i$$

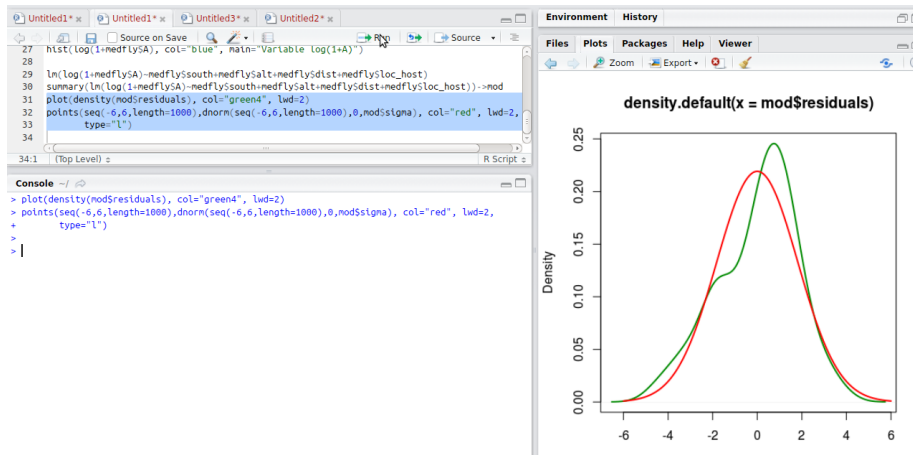
Ejercicio:

- Escribir el modelo para predicción del número de moscas en las trampas que están localizadas en la parte central de Israel.
- Escribir el modelo para predicción del número de moscas en las trampas que están localizadas en la parte sur de Israel.
- Cuál sería su predicción del número de moscas atrapadas en una trampa que está en la parte central, en la altura de 200 metros sobre el nivel del mar, 25 km de la región caliente, y donde no se encuentra ningún huésped dentro de 50 metros de la trampa?

Paso 3: Distribución de los residuos del modelo



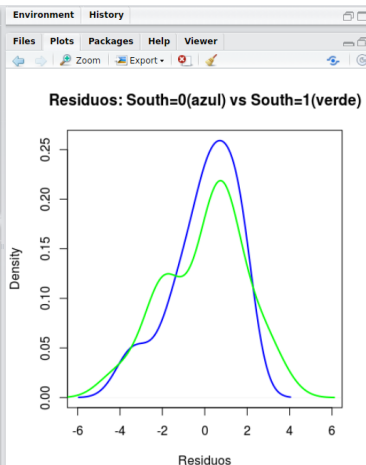
Paso 3: Distribución de los residuos del modelo



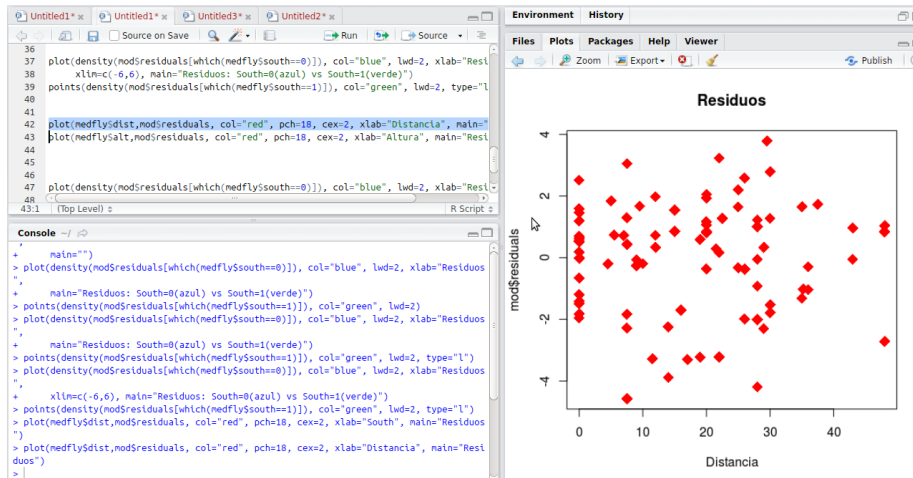
Paso 3: Distribución de los residuos del modelo

```
29 ln(log(1+medfly$A)-medfly$south+medfly$alt+medfly$dist+medfly$loc_host)
30 summary(ln(log(1+medfly$A)-medfly$south+medfly$alt+medfly$dist+medfly$loc_host))>mod
31 plot(density(mod$residuals), col="green4", lwd=2)
32 points(seq(-6,6,length=1000),dnorm(seq(-6,6,length=1000),0,mod$sigma), col="red", lwd=2,
33        type="l")
34
35 plot(medfly$south,mod$residuals, col="red", pch=18, cex=2, xlab="South", main="Residuos")
36
37 plot(density(mod$residuals[which(medfly$south==0)]), col="blue", lwd=2, xlab="Residuos",
38      xlim=c(-6,6), main="Residuos: South=0(azul) vs South=1(verde)")
39 points(density(mod$residuals[which(medfly$south==1)]), col="green", lwd=2, type="l")
40
40:1 (Top Level)
R Script

Console
> plot(density(mod$residuals[which(medfly$south==0)]))
> plot(density(mod$residuals[which(medfly$south==0)]), col="blue")
> plot(density(mod$residuals[which(medfly$south==0)]), col="blue", lwd=2)
> plot(density(mod$residuals[which(medfly$south==0)]), col="blue", lwd=2, xlab="Residuos")
> plot(density(mod$residuals[which(medfly$south==0)]), col="blue", lwd=2, xlab="Residuos")
> plot(density(mod$residuals[which(medfly$south==0)]), col="blue", lwd=2, xlab="Residuos",
+      main="")
> plot(density(mod$residuals[which(medfly$south==0)]), col="blue", lwd=2, xlab="Residuos",
+      main="Residuos: South=0(azul) vs South=1(verde)")
> points(density(mod$residuals[which(medfly$south==1)]), col="green", lwd=2)
> plot(density(mod$residuals[which(medfly$south==0)]), col="blue", lwd=2, xlab="Residuos",
+      main="Residuos: South=0(azul) vs South=1(verde)")
> points(density(mod$residuals[which(medfly$south==1)]), col="green", lwd=2, type="l")
> plot(density(mod$residuals[which(medfly$south==0)]), col="blue", lwd=2, xlab="Residuos",
+      xlim=c(-6,6), main="Residuos: South=0(azul) vs South=1(verde)")
> points(density(mod$residuals[which(medfly$south==1)]), col="green", lwd=2, type="l")
>
```

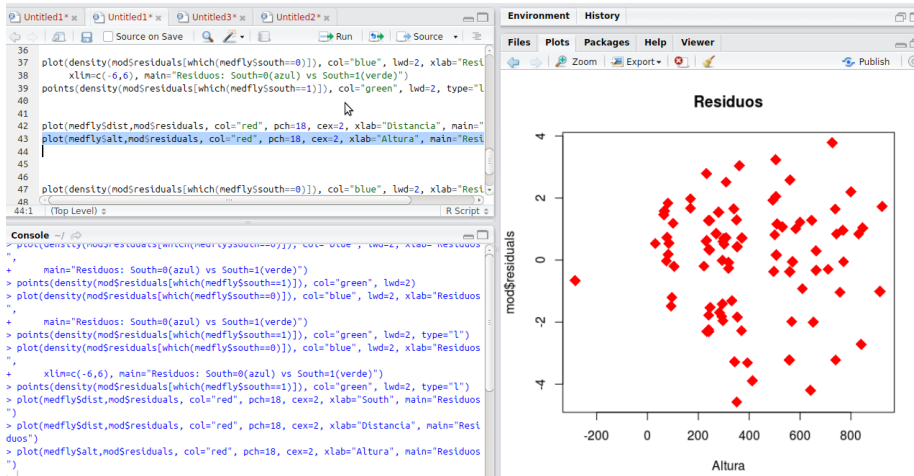


Paso 3: Distribución de los residuos del modelo



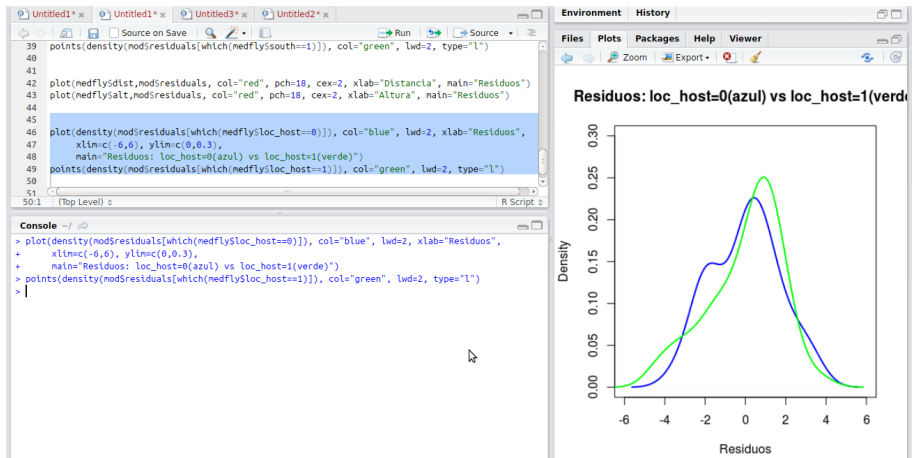
Parece que los valores de los residuos no dependen de la variable distancia.

Paso 3: Distribución de los residuos del modelo



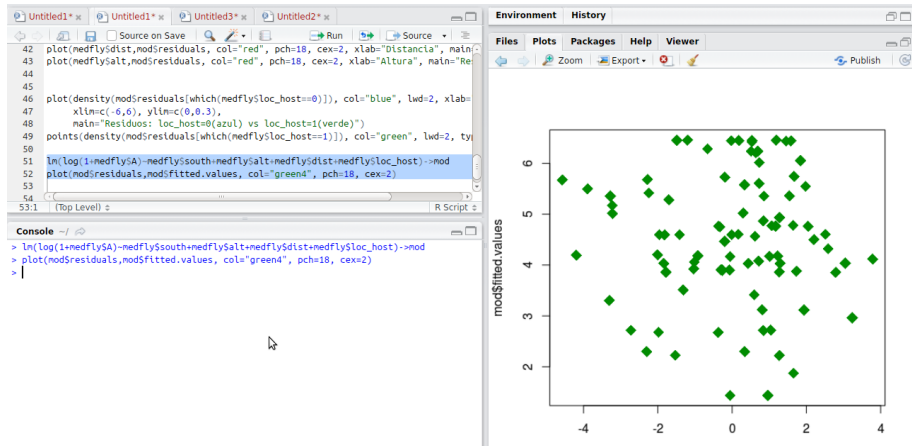
Parece que los valores de los residuos no dependen de la variable altura.

Paso 3: Distribución de los residuos del modelo



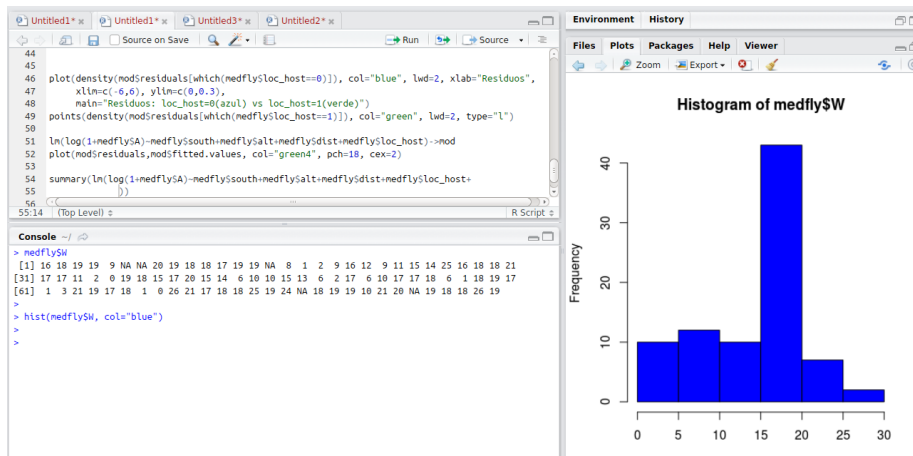
Parece que los valores de los residuos no dependen de la variable loc host.

Paso 3: Residuos vs. Predicciones



Los residuos y las predicciones no están correlacionados

Paso 4: Mejoramiento del Modelo



Paso 4: Mejoramiento del Modelo

```
> which(medfly$W>=0)
[1] 1 2 3 4 5 8 9 10 11 12 13 14 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
[31] 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
[61] 64 65 66 67 68 69 70 71 72 73 74 75 76 78 79 80 81 82 83 85 86 87 88 89
>
> length(which(medfly$W>=0))
[1] 84
>
> which(medfly$W>=0 & medfly$W<=15)
[1] 5 16 17 18 19 21 22 23 24 25 33 34 35 38 41 42 43 44 45 46 47 48 49 51 52 56 57 61 62 67
[31] 68 81
>
> W1=rep(0,89)
> W1[which(medfly$W>=0 & medfly$W<=15)]<-1
>
> W1
[1] 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 1 1 1 0 0 0 0 0 0 0 0 1 1 1 0 0 1 0 0 1 1 1 1 1
[46] 1 1 1 1 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
>
> W2=rep(0,89)
>
> W2[which(medfly$W>15 & medfly$W<=20)]<-1
>
> W2
[1] 1 1 1 1 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 1 0 1 1 0 0 0 1 1 0 1 1 0 0 0 0 0
[46] 0 0 0 0 1 0 0 1 1 1 0 0 1 1 1 0 0 0 1 1 1 0 0 0 0 1 1 1 0 1 0 0 1 1 1 0 0 1 0 1 1 1 0 1
> |
```

Paso 4: Mejoramiento del Modelo

```
> W2
[1] 1 1 1 1 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 1 0 0 0 0 0 0 1 1 1 0 1 1 0 0 0 1 1 0 1 1 0 0 0 0 0
[46] 0 0 0 0 1 0 0 1 1 1 0 0 1 1 1 0 0 0 1 1 1 0 0 0 0 1 1 1 0 1 0 0 1 1 1 0 0 1 0 1 1 1 0 1
> summary(lm(log(1+medfly$A)~medfly$south+medfly$Salt+medfly$dist+medfly$loc_host+W1+W2
+
  ))
```

Call:

```
lm(formula = log(1 + medfly$A) ~ medfly$south + medfly$Salt +
    medfly$dist + medfly$loc_host + W1 + W2)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.5781	-1.0070	0.3347	0.9133	3.2375

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2.6296893	0.8131629	3.234	0.001761	**
medfly\$south	-0.3297513	0.4556742	-0.724	0.471336	
medfly\$Salt	0.0008477	0.0011456	0.740	0.461418	
medfly\$dist	-0.0554868	0.0211881	-2.619	0.010509	*
medfly\$loc_host	1.1477107	0.4668316	2.459	0.016057	*
W1	2.3478226	0.7406785	3.170	0.002145	**
W2	2.2356049	0.5614663	3.982	0.000147	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.685 on 82 degrees of freedom

Multiple R-squared: 0.4539, Adjusted R-squared: 0.4139

F-statistic: 11.36 on 6 and 82 DF, p-value: 3.335e-09

Paso 4: Mejoramiento del Modelo

$$\hat{Y}_i = 2.63 - 0.33S_i + 0.08Alt_i - 0.055Dist_i + 1.14H_i + 2.35W1_i + 2.24W2_i$$

Si la primera mosca fue capturada dentro de 15 primeras semanas, el modelo es:

$$\hat{Y}_i = 4.98 - 0.33S_i + 0.08Alt_i - 0.055Dist_i + 1.14H_i$$

Si la primera mosca fue capturada dentro de 15-20 primeras semanas, el modelo es:

$$\hat{Y}_i = 4.87 - 0.33S_i + 0.08Alt_i - 0.055Dist_i + 1.14H_i$$

Si la primera captura ocurrió después de 20 primeras semana o nunca ocurrió, el modelo es:

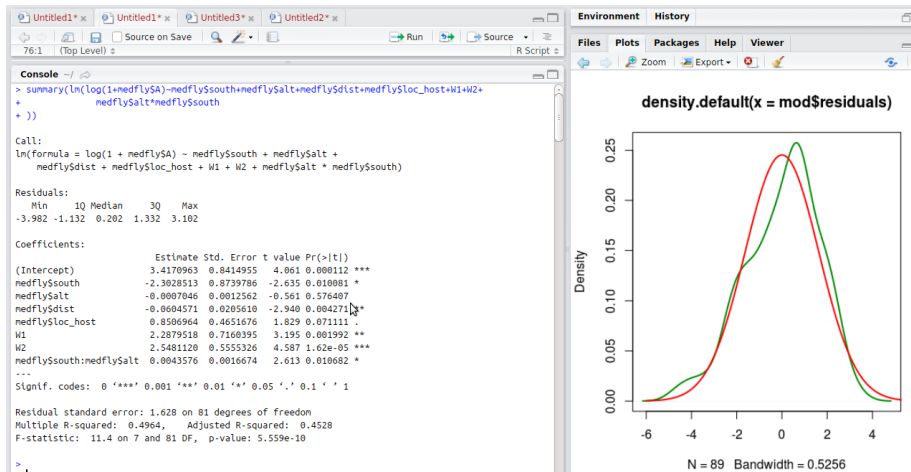
$$\hat{Y}_i = 2.63 - 0.33S_i + 0.08Alt_i - 0.055Dist_i + 1.14H_i$$

Paso 4: Mejoramiento del Modelo

Ejercicio:

- Escribir el modelo para predicción del número de moscas en las trampas que están localizadas en la parte central de Israel, si la primera captura ocurrió en la semana 13.
- Escribir el modelo para predicción del número de moscas en las trampas que están localizadas en la parte sur de Israel, si la primera captura ocurrió en la semana 24.
- Cuál sería su predicción del número de moscas atrapadas en una trampa que esta en la parte central, en la altura de 200 metros sobre el nivel del mar, 25 km de la región caliente, y donde no se encuentra ningun huesped dentro de 50 metros de la trampa, y la primera captura ocurrió en la semana 17?
- Aplicar el modelo con más categorías de la variable W. Explicar los resultados.

Paso 4: Mejoramiento del Modelo: modelo con interacción



Paso 4: Mejoramiento del Modelo: modelo con interacción

$$\hat{Y}_i = 3.42 - 2.30S_i - 0.0007Alt_i - 0.06Dist_i + 0.85H_i + 2.38W1_i + 2.55W2_i + 0.0044S_i * Alt_i$$

Parte Central:

$$\hat{Y}_i = 3.42 - 0.0007Alt_i - 0.06Dist_i + 0.85H_i + 2.38W1_i + 2.55W2_i$$

Parte Sur:

$$\hat{Y}_i = 1.12 + (-0.0007 + 0.0044)Alt_i - 0.06Dist_i + 0.85H_i + 2.38W1_i + 2.55W2_i$$

Ejercicio: Ajustar modelos con otras interacciones. Explicar los resultados.

Ejercicio: Cuál sería su predicción del número de moscas atrapadas en una trampa que esta en la parte central, en la altura de 200 metros sobre el nivel del mar, 25 km de la región caliente, y donde no se encuentra ningun huesped dentro de 50 metros de la trampa, y la primera captura ocurrió en la semana 17?

Ejercicio: Ajustar modelos de regresión lineal utilizando las variables x e y (además de las variables: South, Alt, dist, loc host y W)