Inference on Linear Regression Analysis of Covariance

Eli Gurarie

StatR 101 - Lecture 11b
December 3, 2012

December 3, 2012

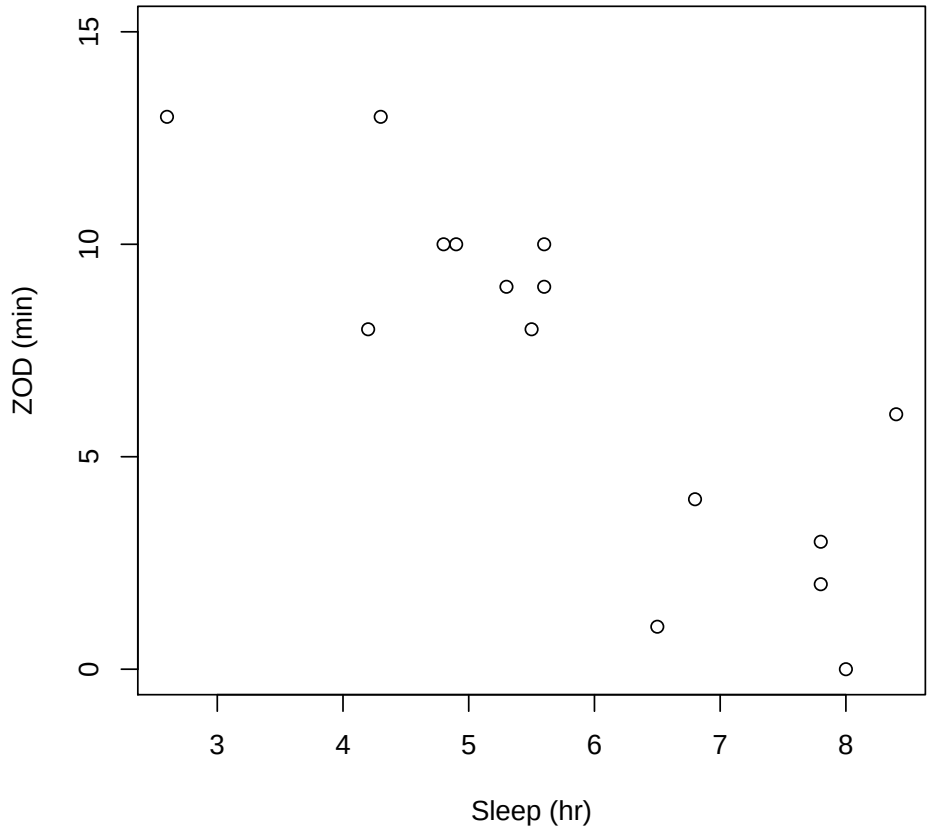
The Stat 311 Insomnia Experiment

During a StatR 101 class, the zone-outs duration (ZOD, in minutes) of fifteen students was carefully recorded by an investigator. Later, the investigator surveyed the students on how many hours the students slept the night before. The results are tabulated below (ordered by amount of sleep):

Student	1	2	3	4	5
Sleep (hours)	2.6	4.2	4.3	4.8	4.9
ZOD (min)	13	8	13	10	10
Student	6	7	8	9	10
S	5.3	5.5	5.6	5.6	6.5
ZOD	9	8	9	10	1
Student	11	12	13	14	15
S	6.8	7.8	7.8	8.0	8.4
ZOD	4	3	2	0	6

Results

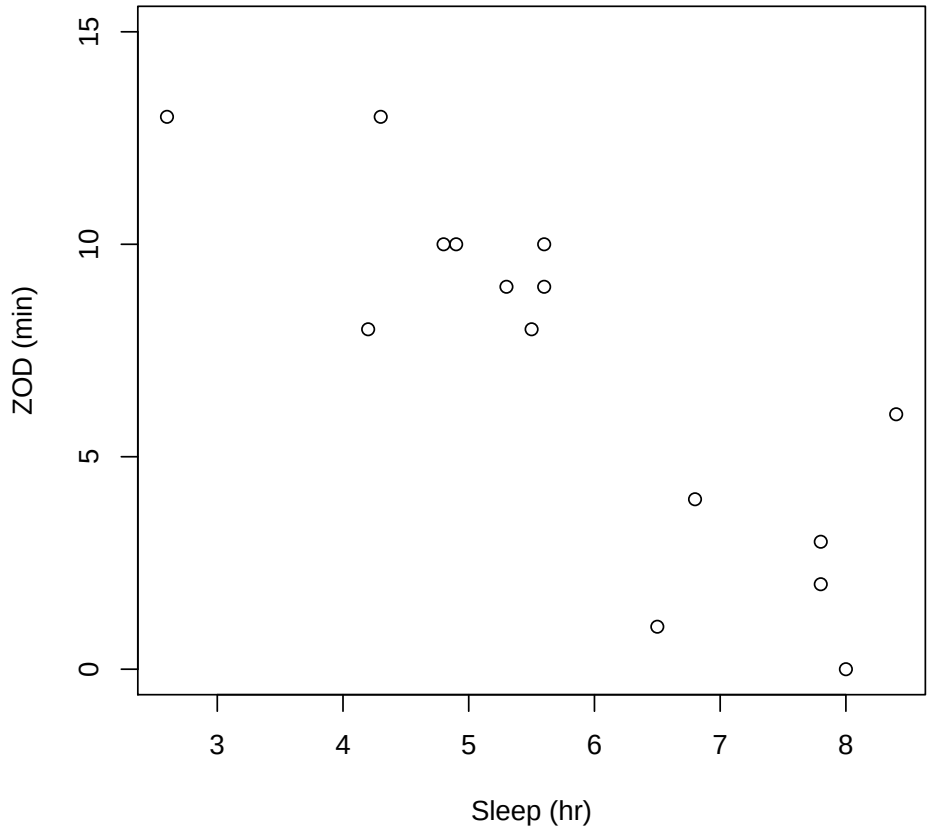
Looks like there might be a relationship!



Perhaps linear?

Results

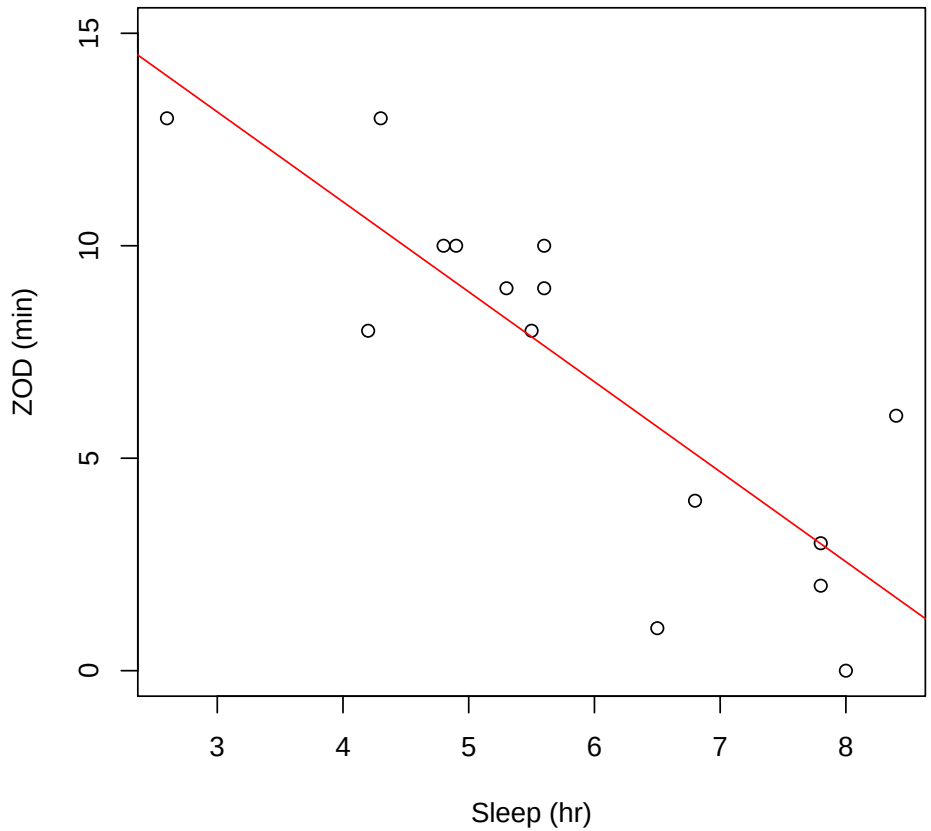
Looks like there might be a relationship!



Perhaps linear?

Results

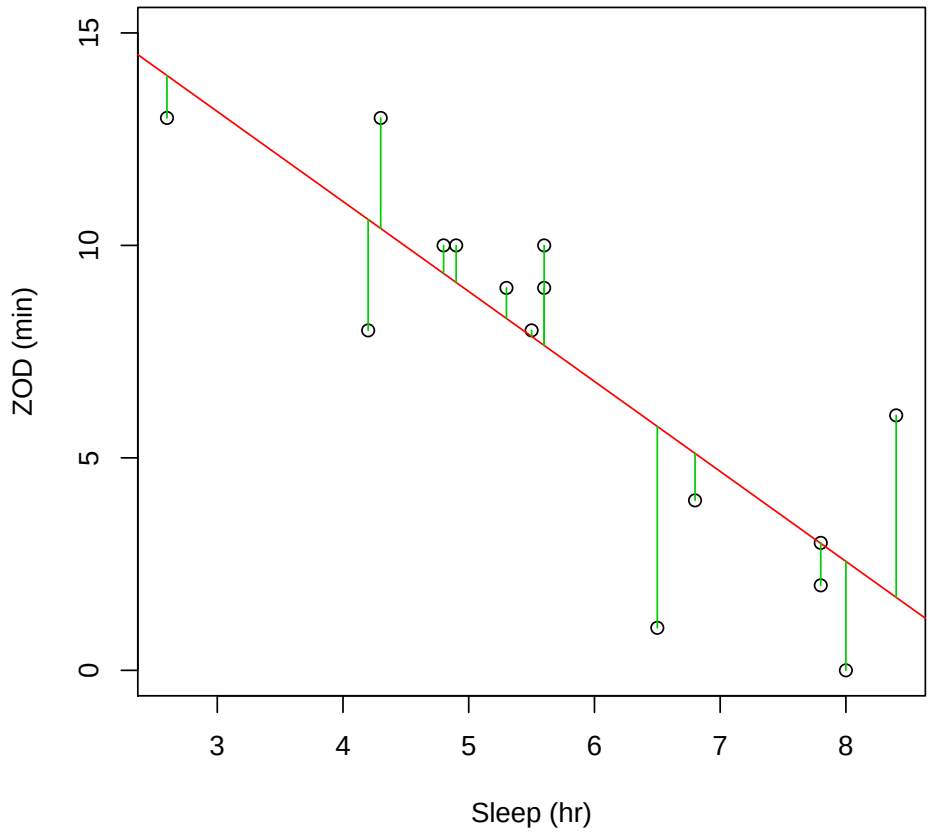
We propose a simple linear model: $Y_i = \alpha + \beta X_i + \epsilon_i$



If $\epsilon_i \sim \text{Normal}(0, \sigma)$ AND ARE iid we can apply the simplest form of *linear regression*.

Method of least squares

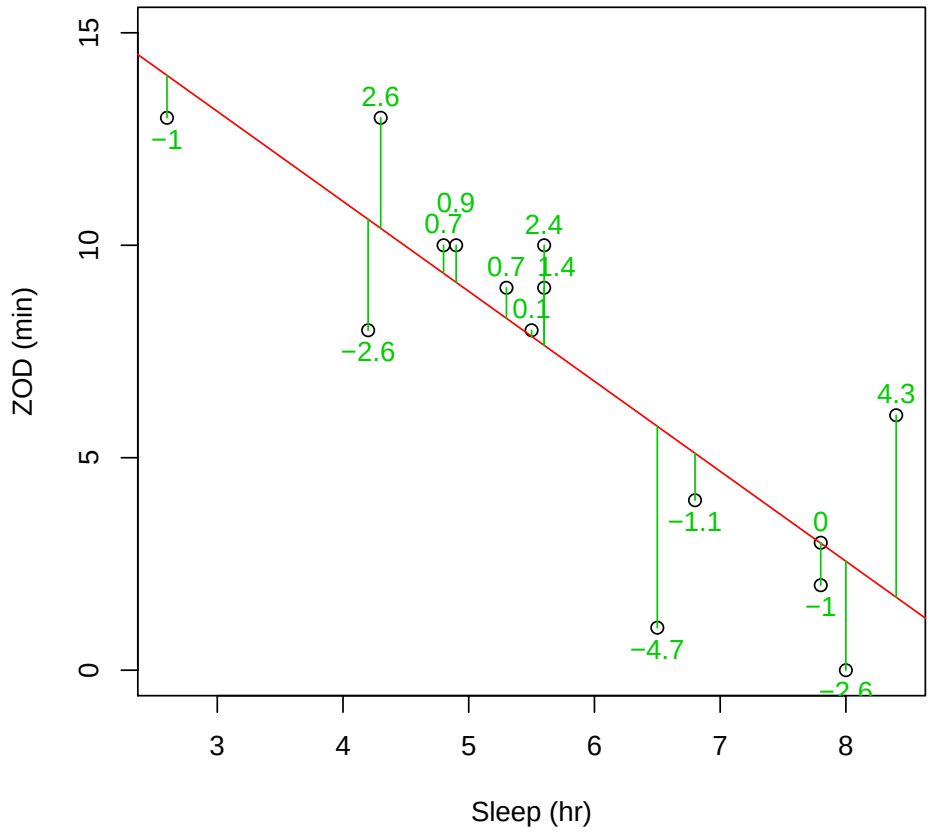
Draw some line (model)!



Find the distances between the points and the line $(Y_i - \widehat{Y}_i)$

Method of least squares

Measure the sum of their squares: $SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$



And find values for $\widehat{\beta}$ and $\widehat{\alpha}$ that minimizes the SS.

How?

With a little math! Minimization means:

$$\frac{\partial SSE}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\sum_{i=1}^{n}(Y_i - (\alpha + \beta X_i))^2\right) = 0$$

$$\frac{\partial SSE}{\partial \alpha} = \frac{\partial}{\partial \alpha}\left(\sum_{i=1}^{n}(Y_i - (\alpha + \beta X_i))^2\right) = 0$$

Solving these two equations yields:

$$\widehat{\beta} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2}$$

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{X}$$

Remember also that:

$$\widehat{\beta} = r_{xy}\frac{s_y}{s_x}$$

Where:

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{Y_i - \overline{Y}}{s_y}\right)\left(\frac{X_i - \overline{X}}{s_x}\right)$$

How?

With a little math! Minimization means:

$$\frac{\partial SSE}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\sum_{i=1}^{n}(Y_i - (\alpha + \beta X_i))^2\right) = 0$$

$$\frac{\partial SSE}{\partial \alpha} = \frac{\partial}{\partial \alpha}\left(\sum_{i=1}^{n}(Y_i - (\alpha + \beta X_i))^2\right) = 0$$

Solving these two equations yields:

$$\widehat{\beta} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2}$$

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{X}$$

Remember also that:

$$\widehat{\beta} = r_{xy}\frac{s_y}{s_x}$$

Where:

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{Y_i - \overline{Y}}{s_y}\right)\left(\frac{X_i - \overline{X}}{s_x}\right)$$

Estimates for sleep example

Plug in our numbers:

$$\begin{aligned} \overline{X} &= 5.87 \\ \overline{Y} &= 7.07 \\ \widehat{\beta} &= -81/38.2 = -2.11 \\ \widehat{\alpha} &= 7.07 + 2.11 * 5.87 = 19.5 \end{aligned}$$

Our minimized SSE is:

$$SSE = \sum (Y_i - (\alpha + \beta X_i))^2 = 73.2$$

Our final model is

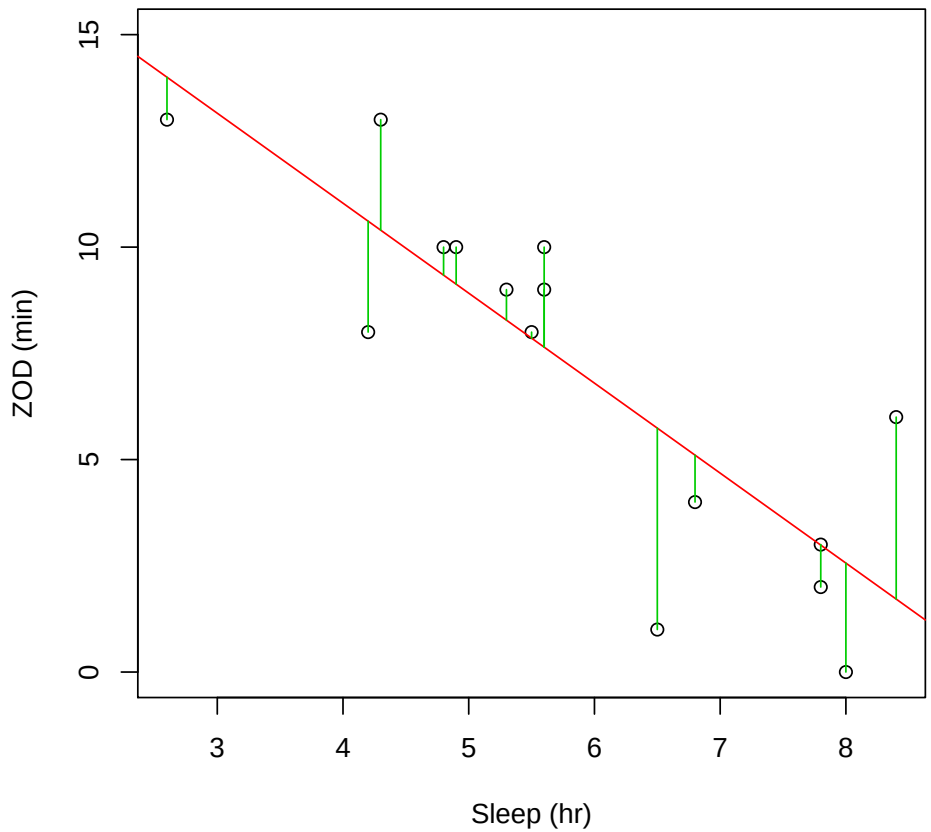
$$Y_i = 19.5 - 2.11 X_i + \epsilon$$

with

$$\widehat{\sigma}^2 = MSE = SSE/(n-2) = 5.64$$

And

Of course, it fits quite nicely:



But how do we assess this model, and do inference?

Some sums of squares

Sum of squares - TOTAL:

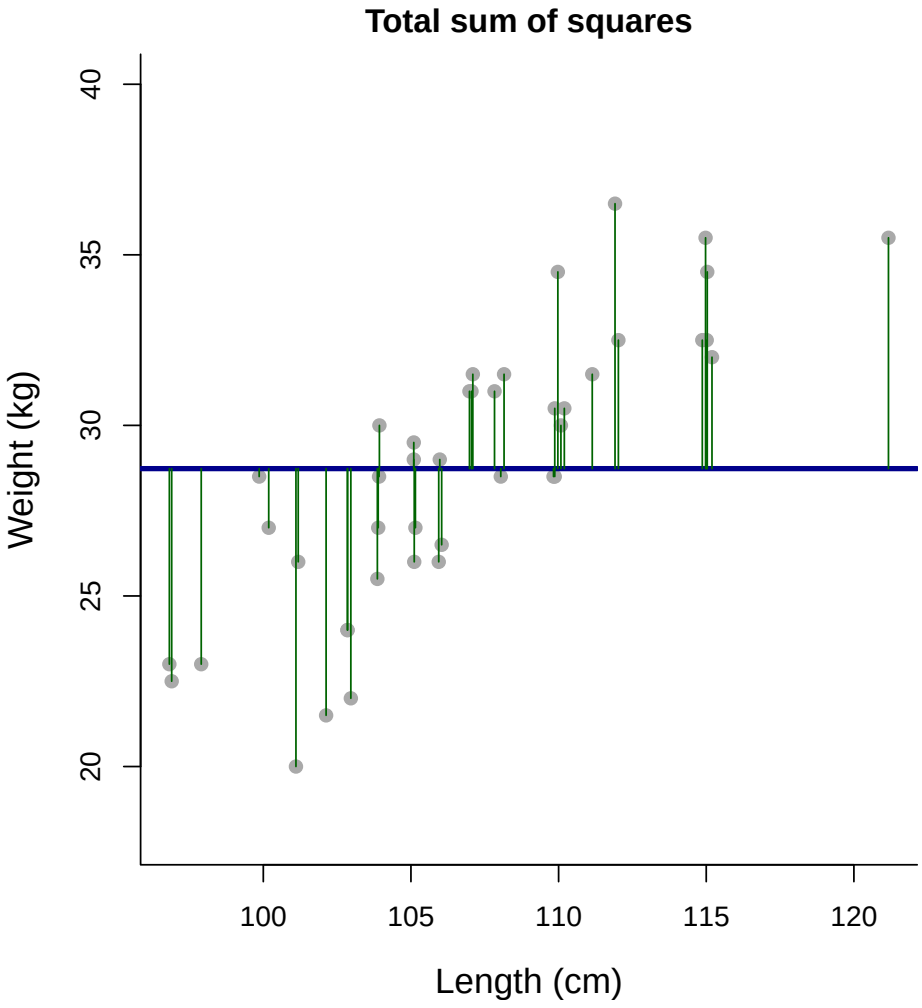
$$SS_{total} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

MODEL sum of squares:

$$SS_{model} = \sum_{i=1}^{n} (\widehat{Y} - \overline{Y})^2$$

ERROR sum of squares:

$$SS_{error} = \sum_{i=1}^{n} (Y_i - \widehat{Y})^2$$



Some sums of squares

Sum of squares - TOTAL:

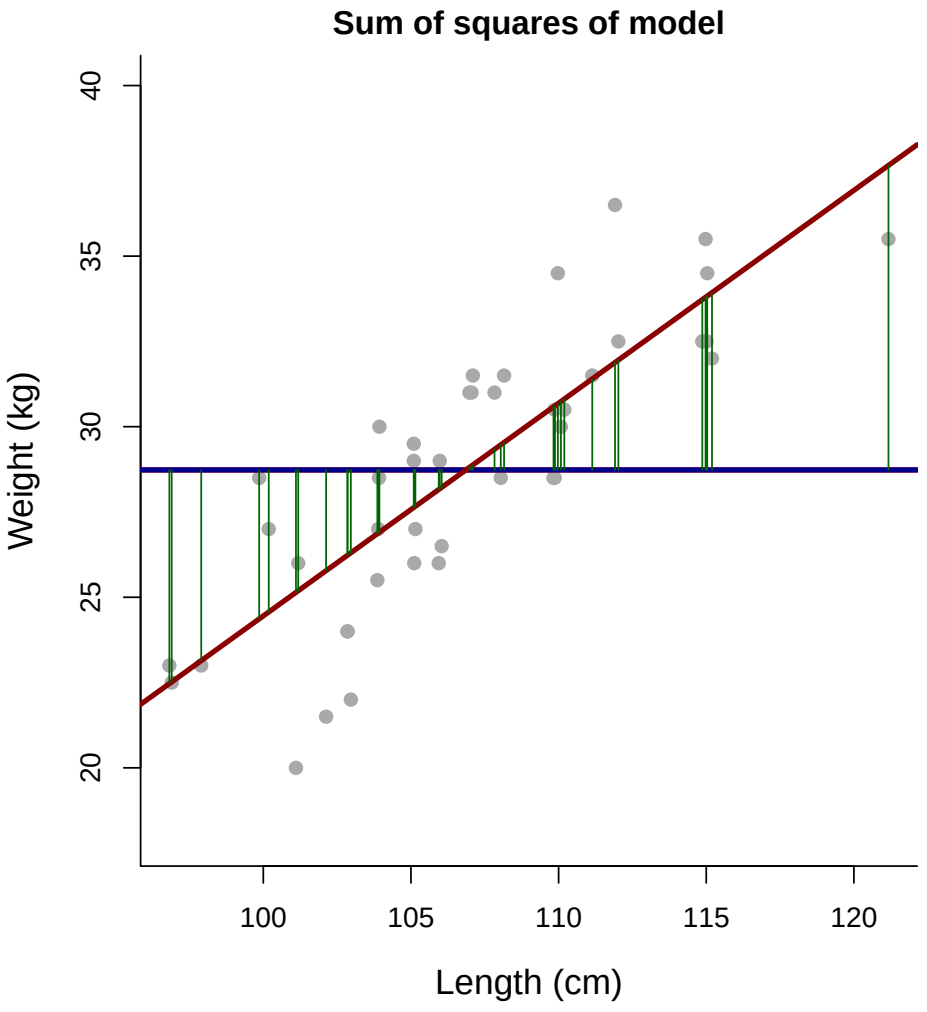
$$SS_{total} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

MODEL sum of squares:

$$SS_{model} = \sum_{i=1}^{n} (\widehat{Y} - \overline{Y})^2$$

ERROR sum of squares:

$$SS_{error} = \sum_{i=1}^{n} (Y_i - \widehat{Y})^2$$



Some sums of squares

Sum of squares - TOTAL:

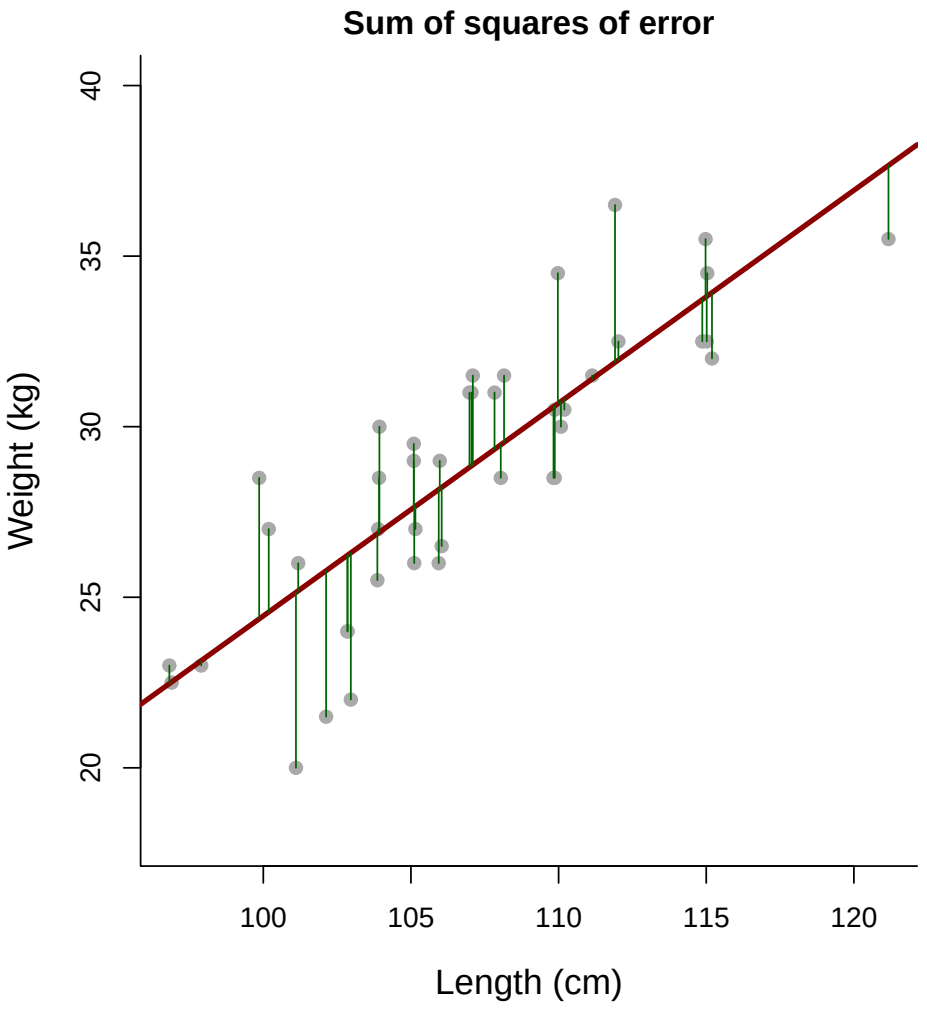
$$SS_{total} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

MODEL sum of squares:

$$SS_{model} = \sum_{i=1}^{n} (\widehat{Y} - \overline{Y})^2$$

ERROR sum of squares:

$$SS_{error} = \sum_{i=1}^{n} (Y_i - \widehat{Y})^2$$



Decomposing the total variance

- We **decompose** the total variation into "explained" and "unexplained" components. So:

$$\text{Total sum of squares} = \text{Regression sum of squares} + \text{Error sum of squares}$$

$$SS_{total} = SS_{model} + SS_{error}$$

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\widehat{Y} - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y})^2$$

Defining the Mean Squares

$$
\begin{aligned}
MS_{total} &= \frac{SS_{total}}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \\
MS_{model} &= SS_{model} = \sum_{i=1}^{n}(\widehat{Y} - \overline{Y})^2 \\
MS_{error} &= \frac{SS_{error}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y})^2
\end{aligned}
$$

Under H_0, each of these is an unbiased estimate of σ^2.

Test statistic

To test:

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

Use the test statistic:

$$F_0 = \frac{MS_{model}}{MS_{error}} = \frac{SS_{model}/1}{SS_{error}/(n-2)}$$

and

$$F_0 \sim \mathcal{F}_{1,n-2}$$

so:

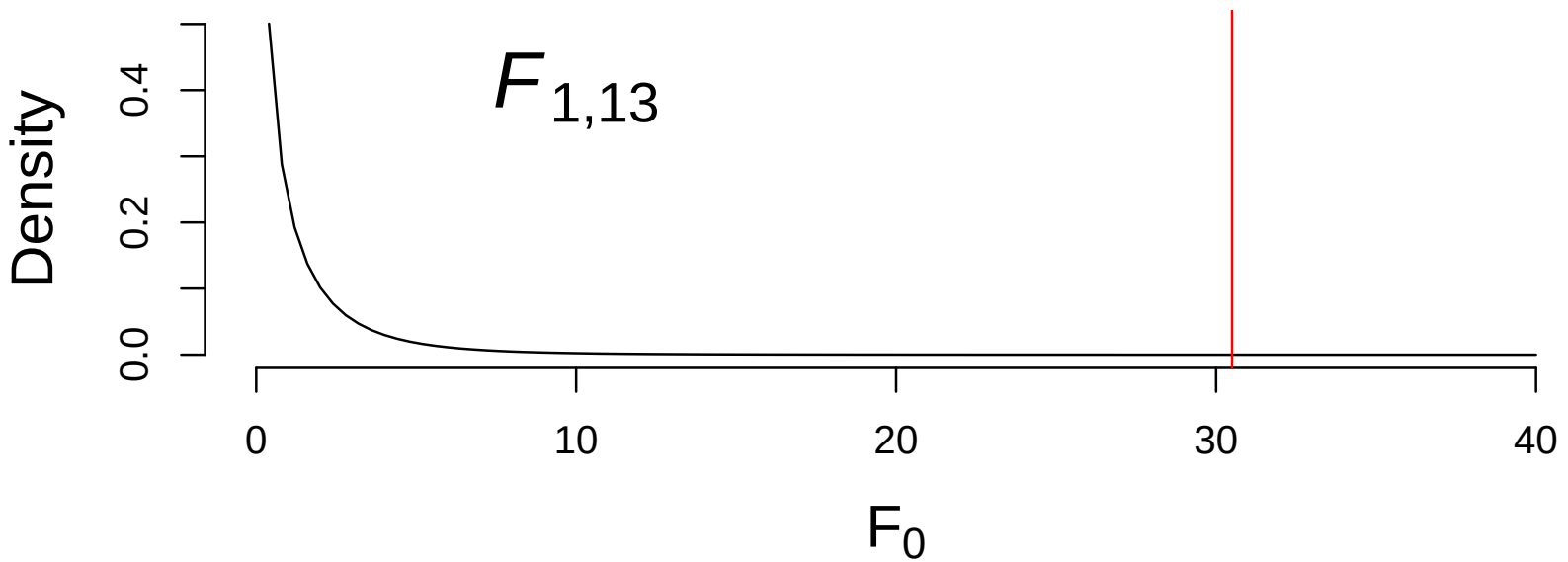
$$p = \Pr\left(\mathcal{F}_{1,n-2} > F_0\right)$$

Anova table

Source of Variation	**Sum of Squares**	**Degrees of freedom**	**Mean Square**	F_0	p-value
Model	SS_{model}	1	MSM	$\frac{MSM}{MSE}$	$Pr[\mathcal{F}_{1,n-2} > F_0]$
Error	SS_{error}	$n-2$	MSE		
Total	SS_{total}	$n-1$			

Anova table: Sleep data

Source of Variation	**Sum of Squares**	**Degrees of freedom**	**Mean Square**	F_0	p-value
Sleep	171.66	1	171.66	30.5	1×10^{-5}
Error	73.27	13	5.636		
Total	244.93	14			



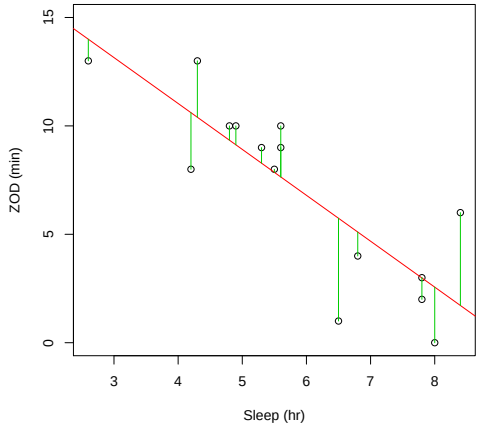
Anova table: Sleep data

Source of Variation	**Sum of Squares**	**Degrees of freedom**	**Mean Square**	F_0	p-value
Sleep	171.66	1	171.66	30.5	1×10^{-5}
Error	73.27	13	5.636		
Total	244.93	14			

1. REJECT THE NULL HYPOTHESIS;
2. conclude that the SLOPE is NOT equal to 0;
3. conclude that there IS an effect of Sleep amount on Zone-Out Duration;
4. collected a small but convincing piece of evidence for a grand theory of statistics student concentration ability.

Model specification

Model:



$$Y = \widehat{\alpha} + \widehat{\beta}X + \epsilon$$

Where:

$$\epsilon \sim N(0, \sigma)$$

Now that we have *chosen* a model, we can *specify* it. Our model has 3 parameters: α, β and σ^2. The estimated values for these parameters are:

parameter	estimate	value
α	$\widehat{\alpha}$	19.5
β	$\widehat{\beta}$	-2.11
σ^2	MS_{error}	5.64

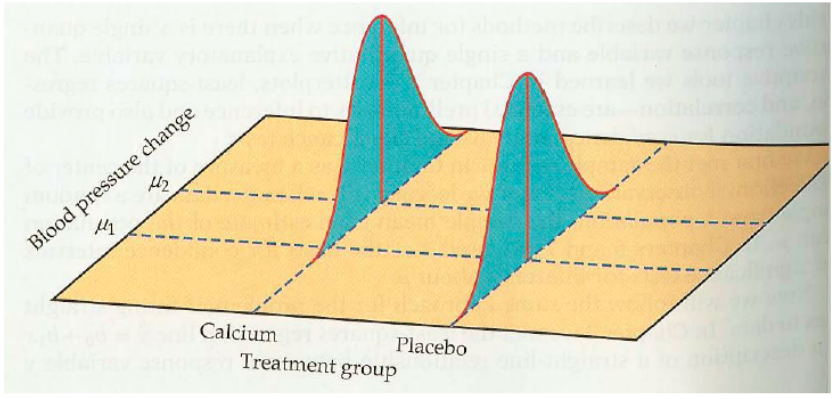
ANOVA for multiple groups and linear regression

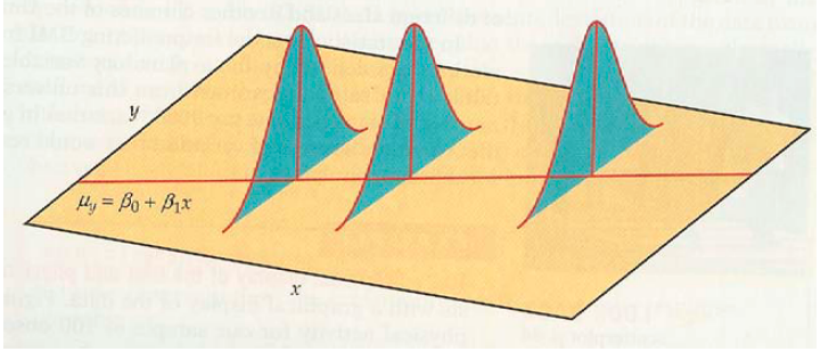
Question:	Are μ_1, μ_2, ... μ_n equal?	Is there a linear relationship: $Y = \alpha + \beta X$?
Test:	One way ANOVA	
Statistics:	$\overline{X_1}, \overline{X_2}, ..., \overline{X_a}$, $SS_{group}, SS_{error}, a \times n = N$	$\overline{Y}, \overline{X}, \widehat{\beta}, \widehat{\alpha}, SS_{error}, SS_{model}, n$
Assumptions:	ϵ all normal, iid (equal variance!)	
H_0:	$\mu_1 = \mu_2 = ... = \mu_n$	$\beta = 0$
H_A:	$\mu_i \neq \mu_j$ for some i and j	$\beta \neq 0$
Test statistic:	$F_0 = \frac{SS_{group}/a-1}{SS_{error}/N-a} = \frac{MSG}{MSE}$	$F_0 = \frac{SS_{model}}{SS_{error}/(n-2)} = \frac{MSM}{MSE}$
Distribution:	$\mathcal{F}_{a-1,N-a}$	$\mathcal{F}_{1,n-2}$
P-value:	$P(\mathcal{F}_{a-1,N-a} > F_0)$	$P(\mathcal{F}_{1,N-2} > F_0)$

Some comments

- Because the actual distributions of $\widehat{\alpha}$ and $\widehat{\beta}$ is Normal, it is possible to construct T statistics to test their hypotheses.
- We did not discuss inference on the intercept parameter α. However, this is usually a parameter of lesser interest and it's distribution is straightforward to derive from the distribution of $\widehat{\beta}$.
- ANOVA on linear regression looks very similar to ANOVA for multiple groups - they are special cases of a general theory.

Comparing inference on Multiple groups and Linear Regression

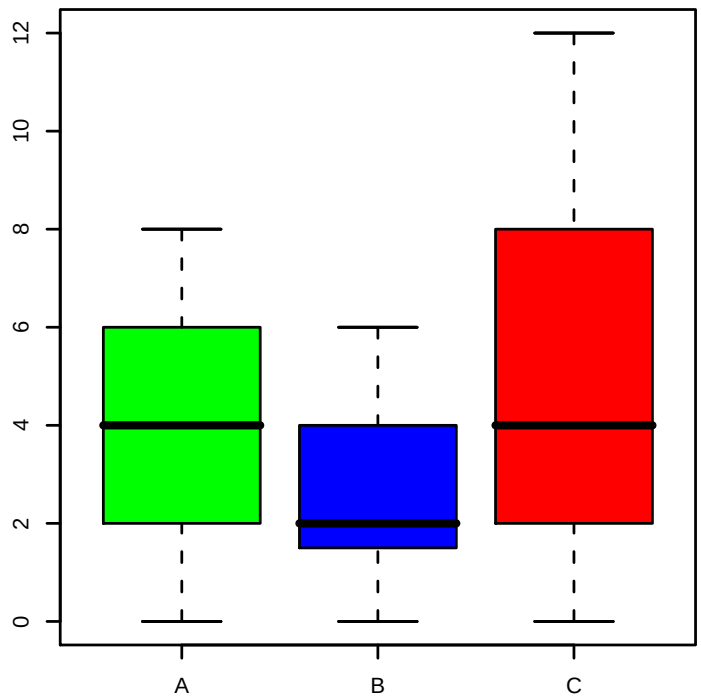


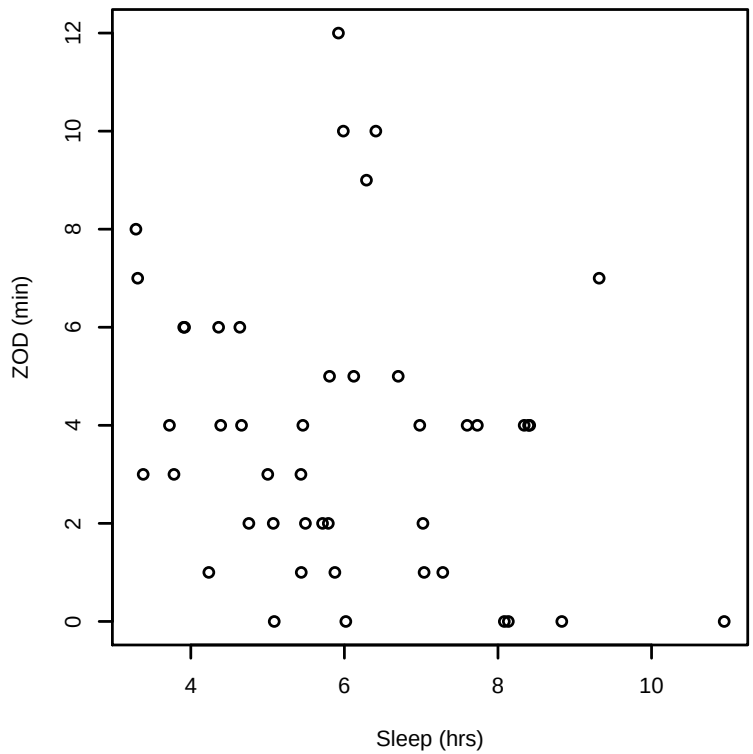


- Group comparison ANOVA
  - continuous response
  - discrete predictor
  - (a-1) parameters for factor level df
- Linear regression ANOVA
  - continuous response
  - continuous predictor
  - 1 parameter summary of association

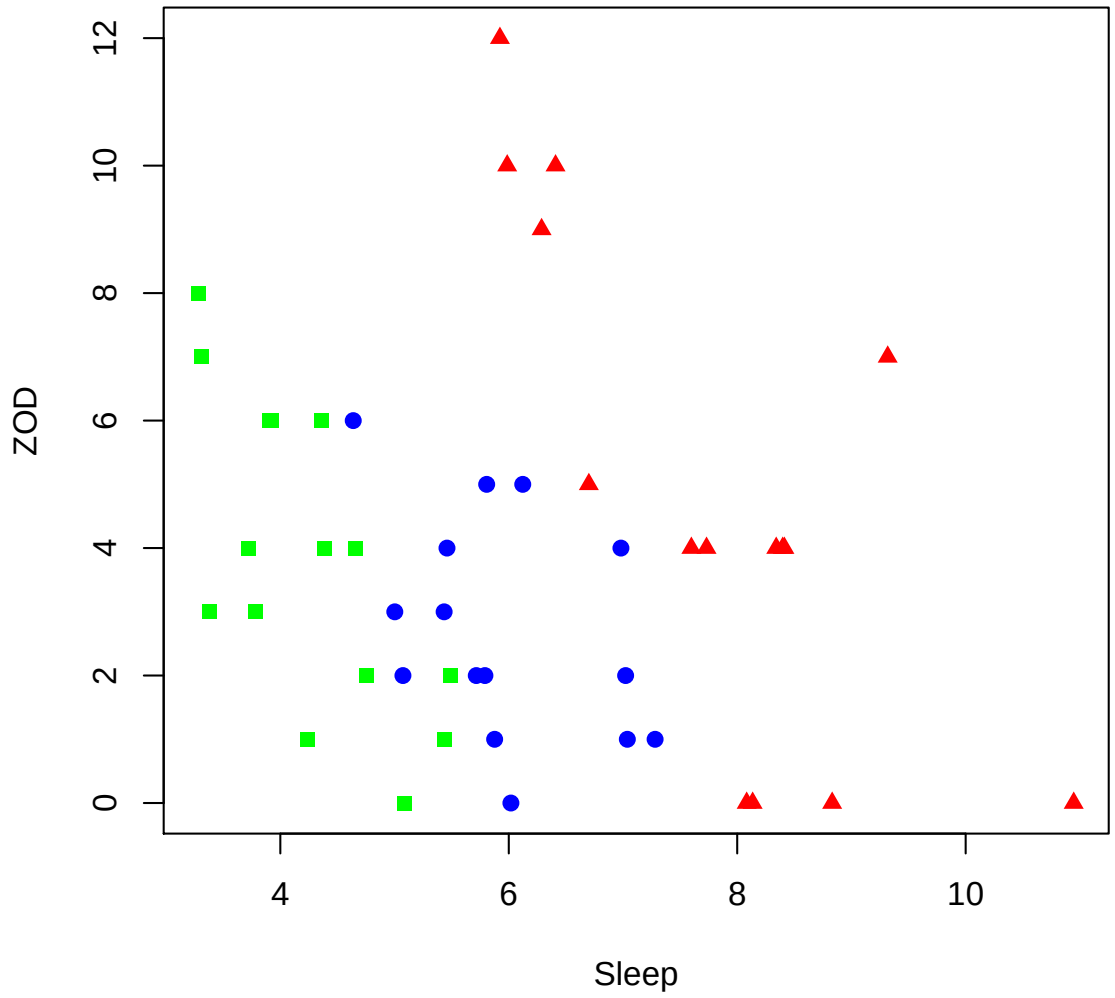
The Great Stat Student Concentration / Pie and Sleep Experiment

In order to integrate the conclusions of the previous two experiments, a researcher has decided to perform the following experiment: over three week, each of fifteen Stat 311 students are randomly assigned one of three pies (apple, blueberryand cherry) to eat and the subsequent ZOD during a linear-regression lecture is monitored. Previous nights sleep data are also collected. The results are plotted below:





A more informative plot:



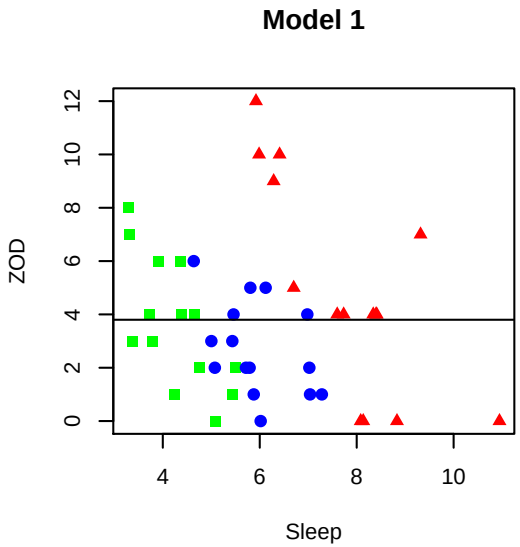
Models

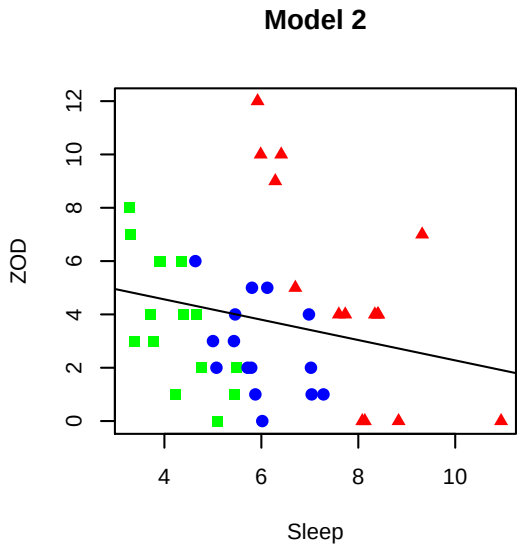
Consider the following models:

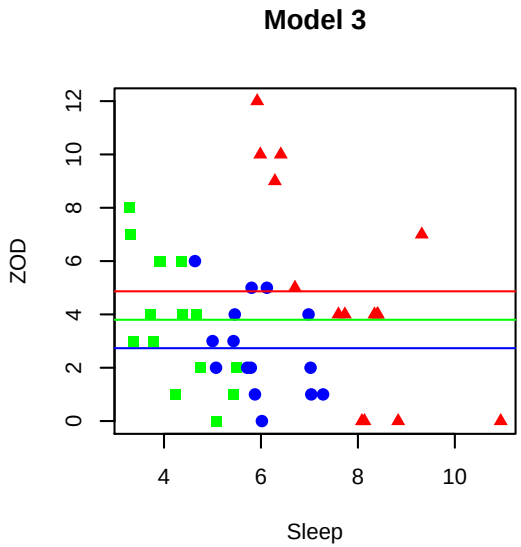
$$
\begin{aligned}
&1: \; Y_i = \mu + \epsilon_i && \text{where } i \in \{1, 2, ...N\} \\
&2: \; Y_i = \alpha + \beta X_i + \epsilon_i && \\
&3: \; Y_{ij} = \mu_i + \epsilon_{ij} && \text{where } i \in \{1, 2, ..., a\} \\
&4: \; Y_{ij} = \alpha_i + \beta X_{ij} + \epsilon_{ij} && \text{and } j \in \{1, 2, ..., n\} \\
&5: \; Y_{ij} = \alpha_i + \beta_i X_{ij} + \epsilon_{ij} &&
\end{aligned}
$$

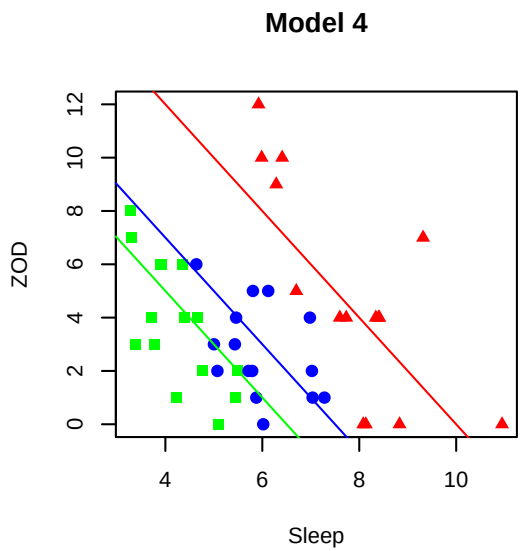
where N is the total number of observations, a is the number of groups, n is the number of measurements per group, μ is the total mean, μ_i is a group mean, α is an intercept parameter, and β is a slope parameter.

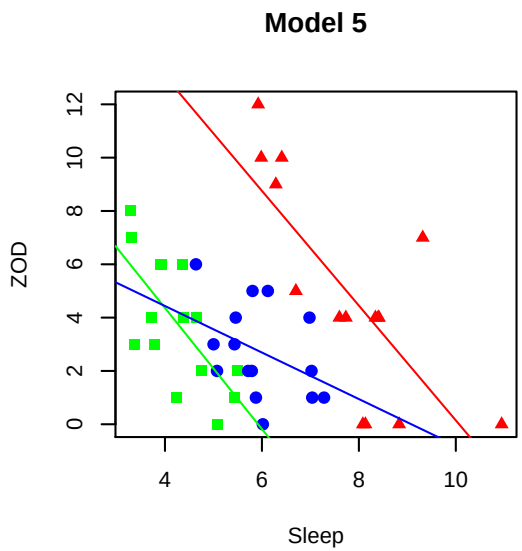
Models


Model 1
ZOD
Sleep
0
2
4
6
8
10
12
4
6
8
10


Model 2
ZOD
Sleep
0
2
4
6
8
10
12
4
6
8
10


Model 3
ZOD
Sleep
0
2
4
6
8
10
12
4
6
8
10


Model 4
ZOD
Sleep
0
2
4
6
8
10
12
4
6
8
10


Model 5
ZOD
Sleep
0
2
4
6
8
10
12
4
6
8
10

Analysis

Without going into any details, here is the ANOVA table for the highest-level model:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Sleep	1	20.14	20.14	4.42	0.0421
Pie	2	170.83	85.41	18.74	0.0000
Sleep:Pie	2	12.45	6.22	1.36	0.2673
Residuals	39	177.79	4.56		

Sleep appears to be a significant factor, Pie appears to be a significant factor, Sleep:Pie (called an *interaction* term), does not appear to be a significant factor.

Based on this table, which of the 5 models is most appropriate? What are its parameter values? How do we interpret and report the results?

Real Example: Coffee, Quality or Certification?

ELSEVIER
www.elsevier.com/locate/worlddev

World Development Vol. 33, No. 3, pp. 497–511, 2005

Printed in Great Britain
0305-750X/$ - see front matter

doi:10.1016/j.worlddev.2004.10.002

Confronting the Coffee Crisis: Can Fair Trade, Organic, and Specialty Coffees Reduce Small-Scale Farmer Vulnerability in Northern Nicaragua?

CHRISTOPHER BACON *
University of California, Santa Cruz, USA

Real Example: Coffee, Quality or Certification?

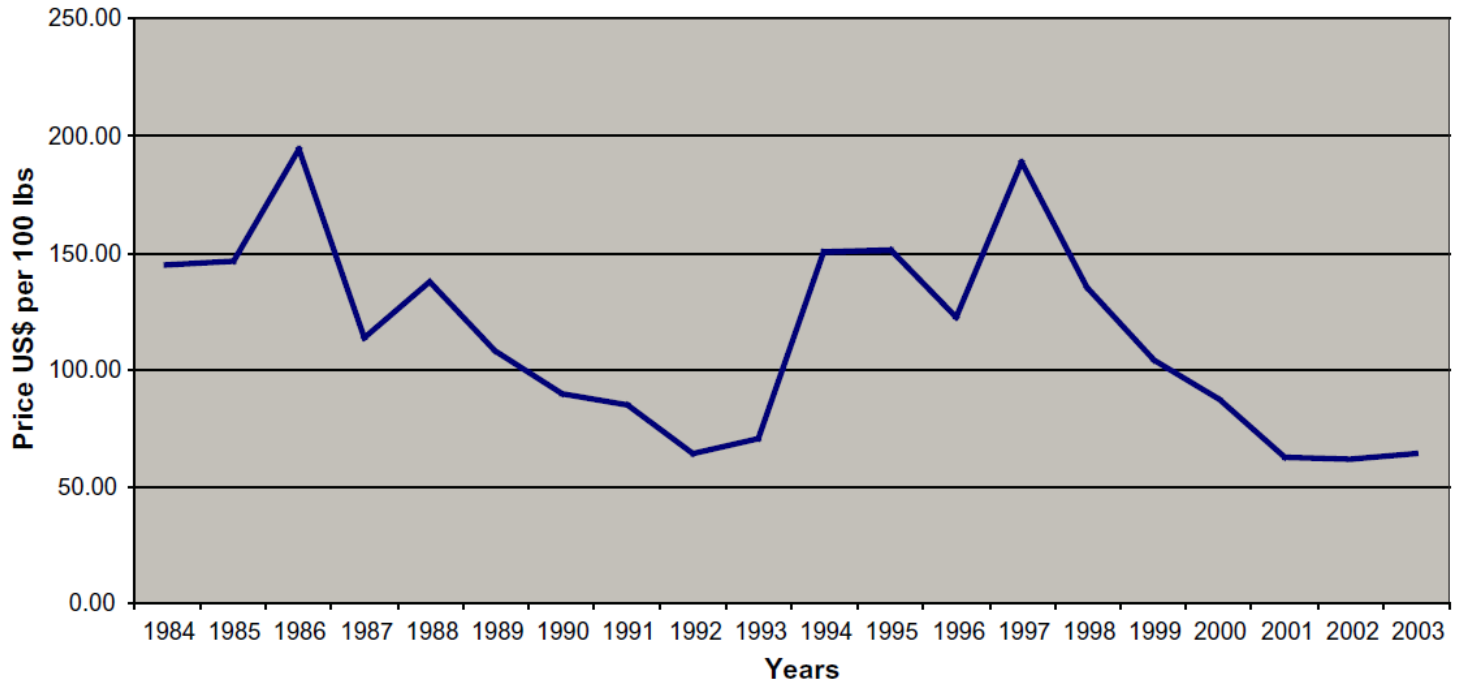
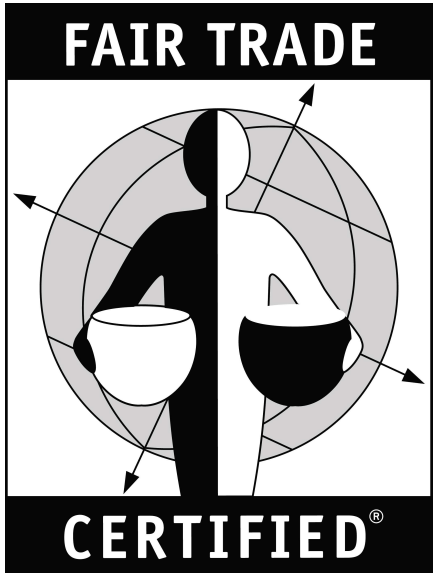


Figure 1. *International coffee prices. Sources: Average yearly prices for arabica coffee beans (other milds) from International Coffee Organization (2003).*

At the time of publication, coffee prices were low, and farmers were paid very little.

Real Example: Coffee, Quality or Certification?



- Better coffee is grown at high altitude
- Some coffees are certified as fair trade, organic, or "bird-friendly" ... others aren't.
- What determines prices paid to farmer? Quality or Certification?

Real Example: Coffee, Quality or Certification?

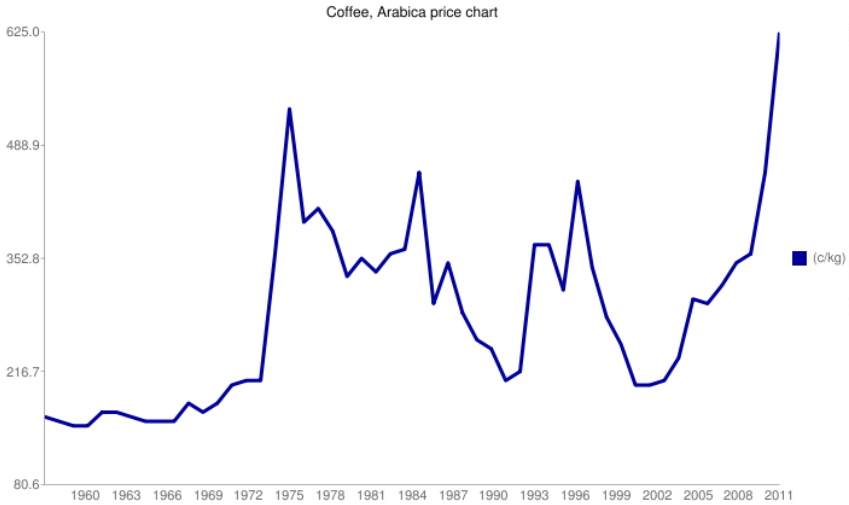
Table 2. *ANOVA results comparing altitude and certification with price as dependent variable*[a]

	DF	Sum of squares	Mean square	*F*-value	*P*-value
Certification	1	1640169.310	1640169.310	78.945	<0.0001
Altitude	1	21131.332	21131.332	1.017	0.3144
Certification & altitude	1	34775.200	34775.200	1.674	0.1972
Residual	209	4342184.525	20776.003		

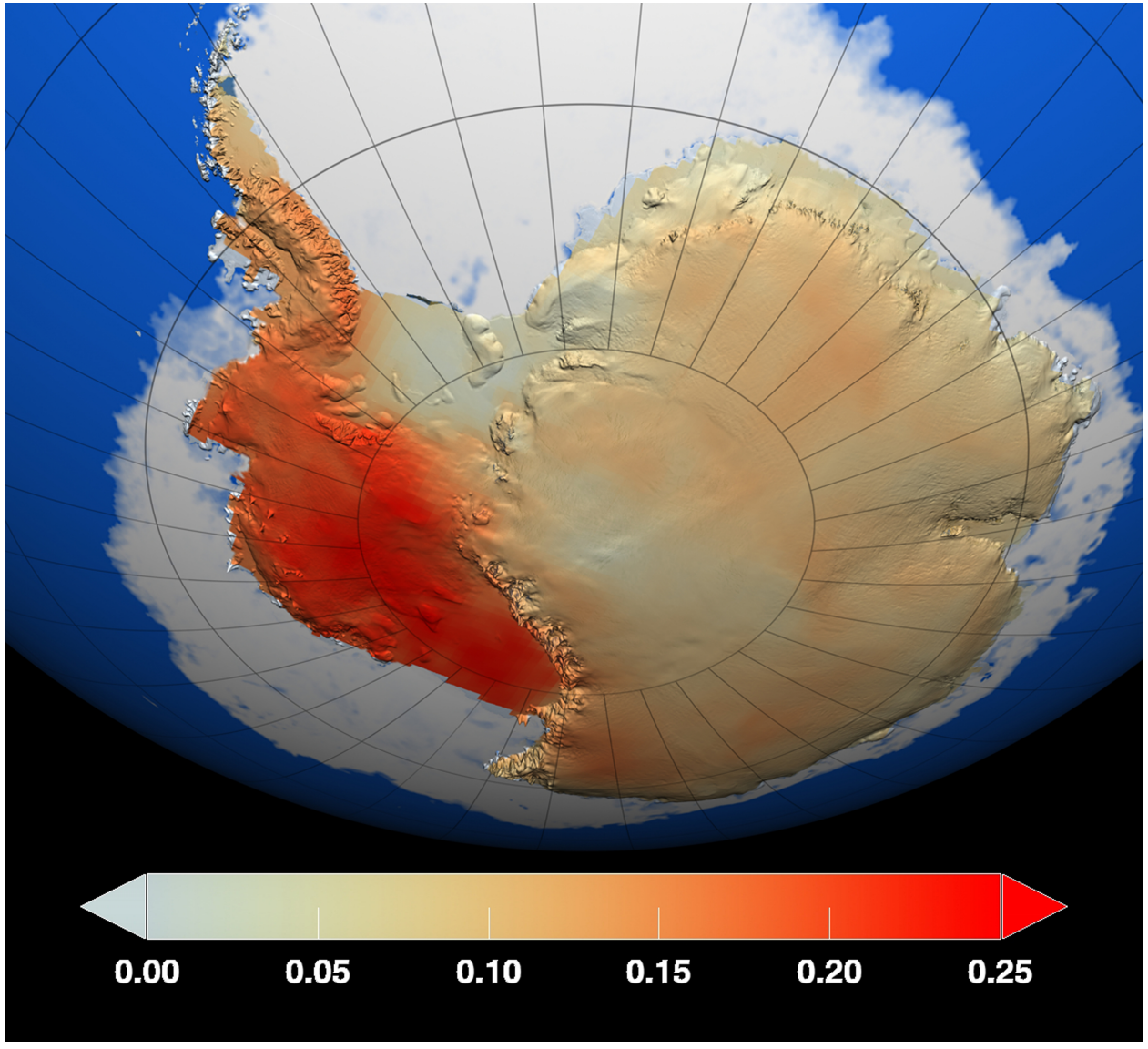
Source: Participatory survey 2001.

Conclusion: "A systematic comparison of price and altitude reveals a **statistically insignificant correlation** between **altitude and price**. I used the average prices in local currency to run **a two-way ANOVA** comparing the impact of altitude and certification on coffee prices... **Certification has a greater influence on price than altitude (quality)**.

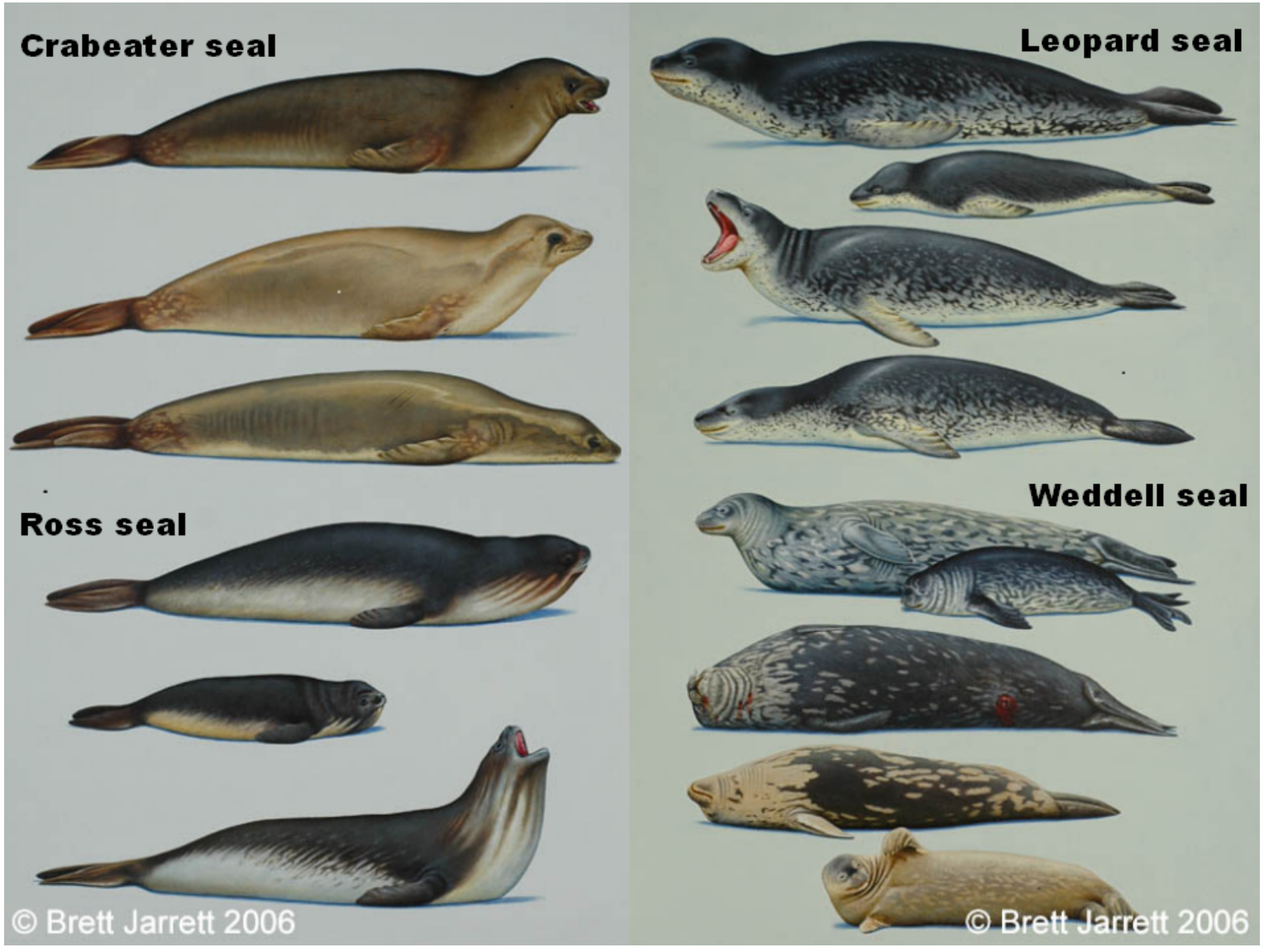
Coffee prices



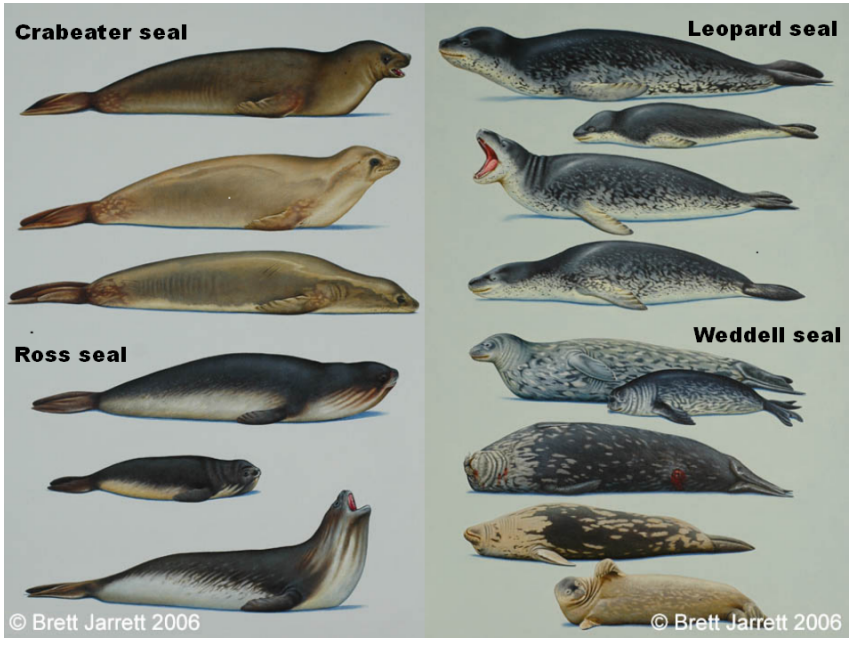
Real example II: Antarctica:



Real example II: Antarctic seals

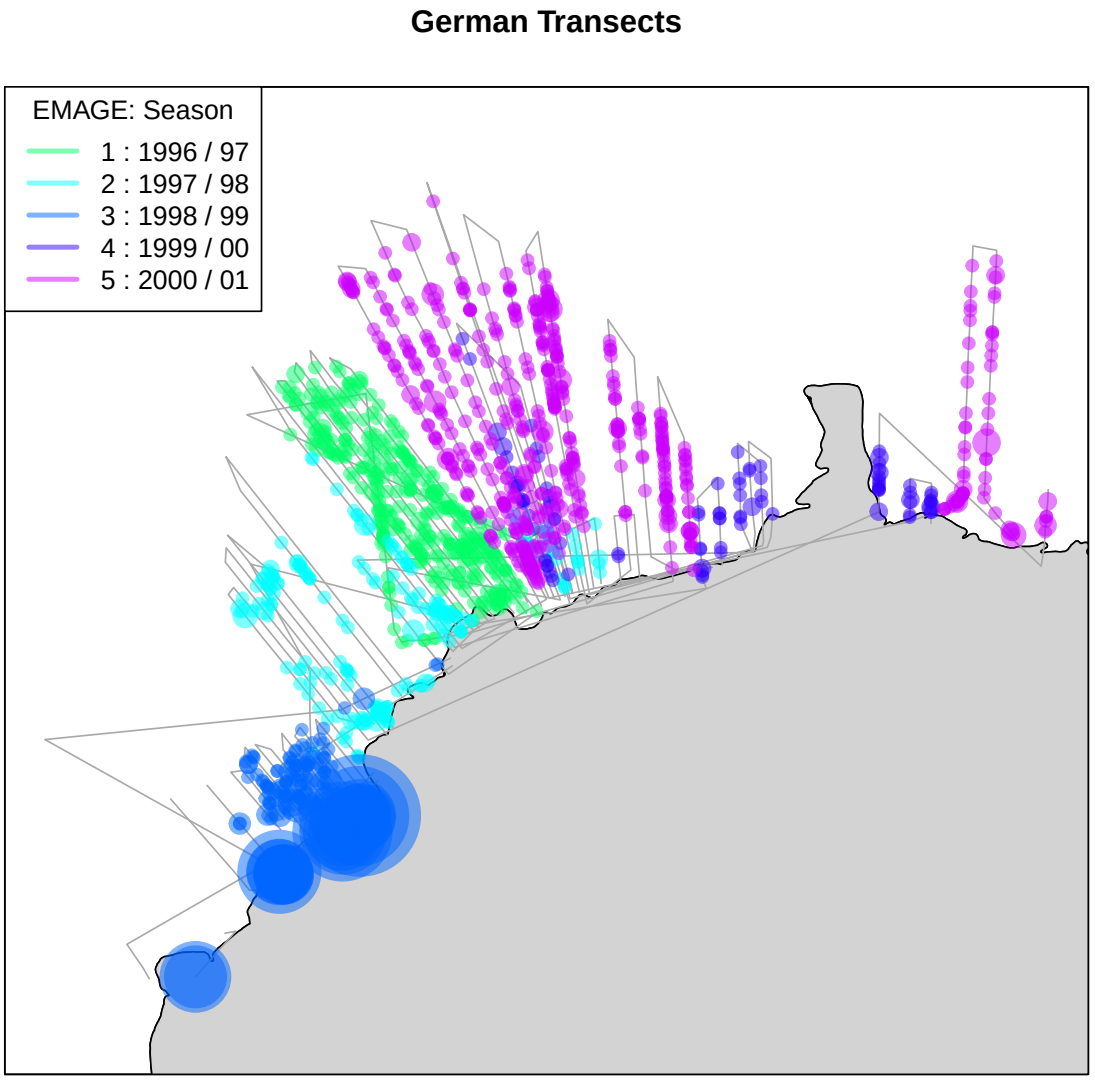


Some questions:



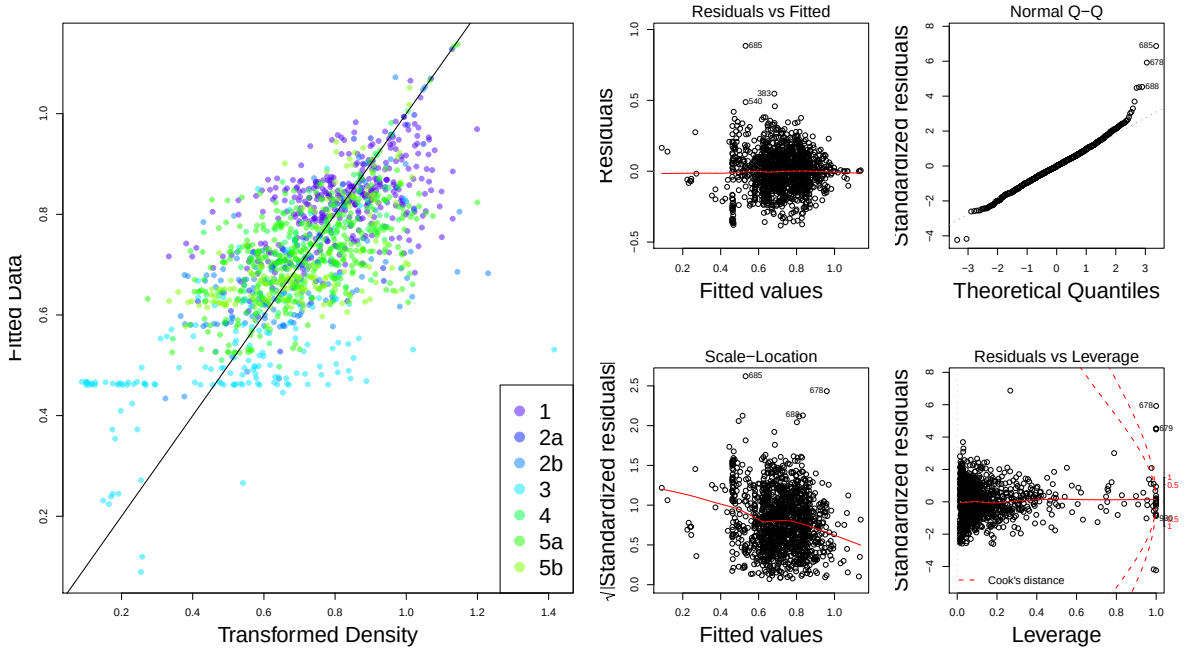
- How many seals are there in Antarctica?
- Are there trends in their population?
- How are seals affected by changes in temperature/ice-cover?

Here is the sample:

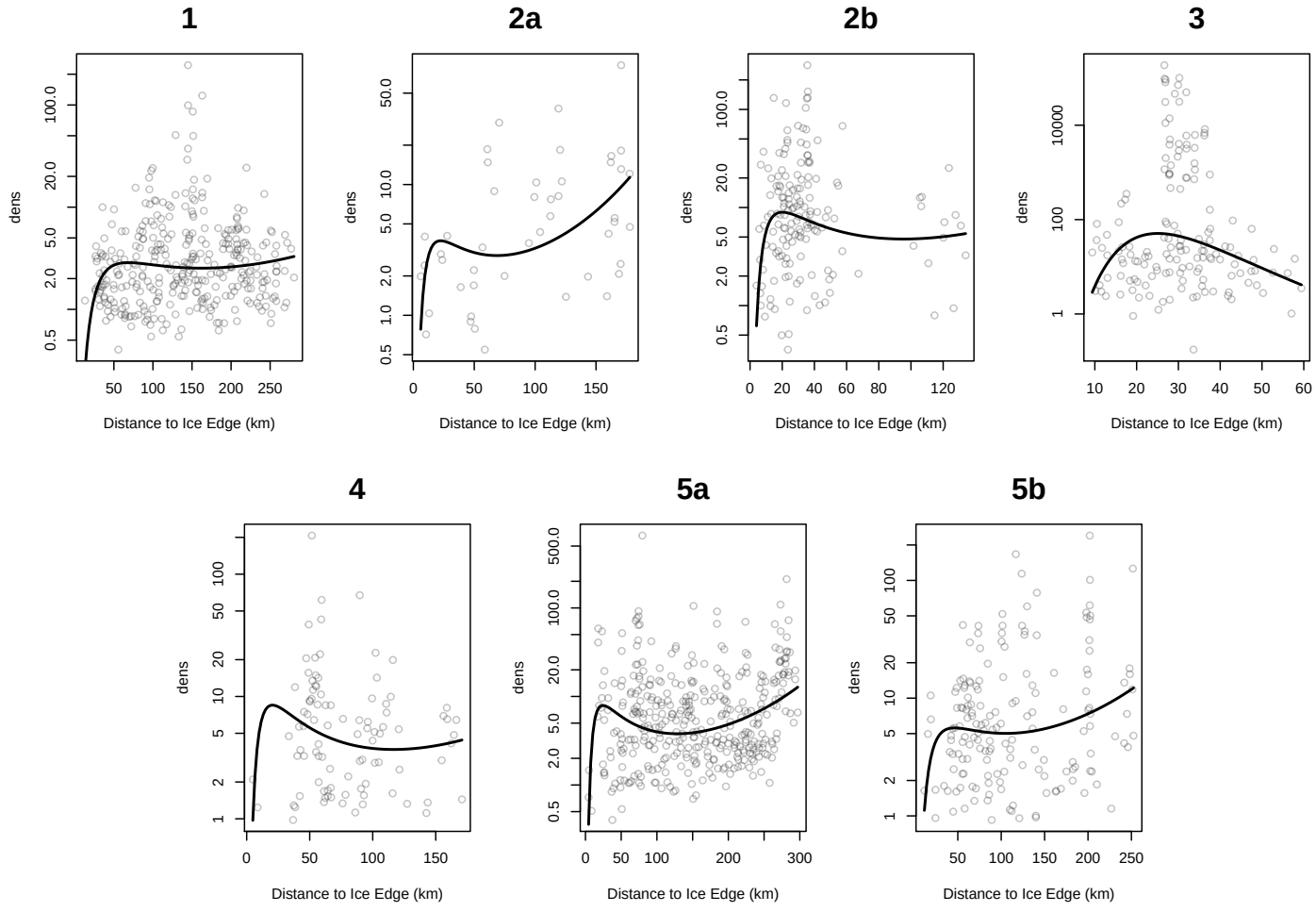


Complicated statistical model

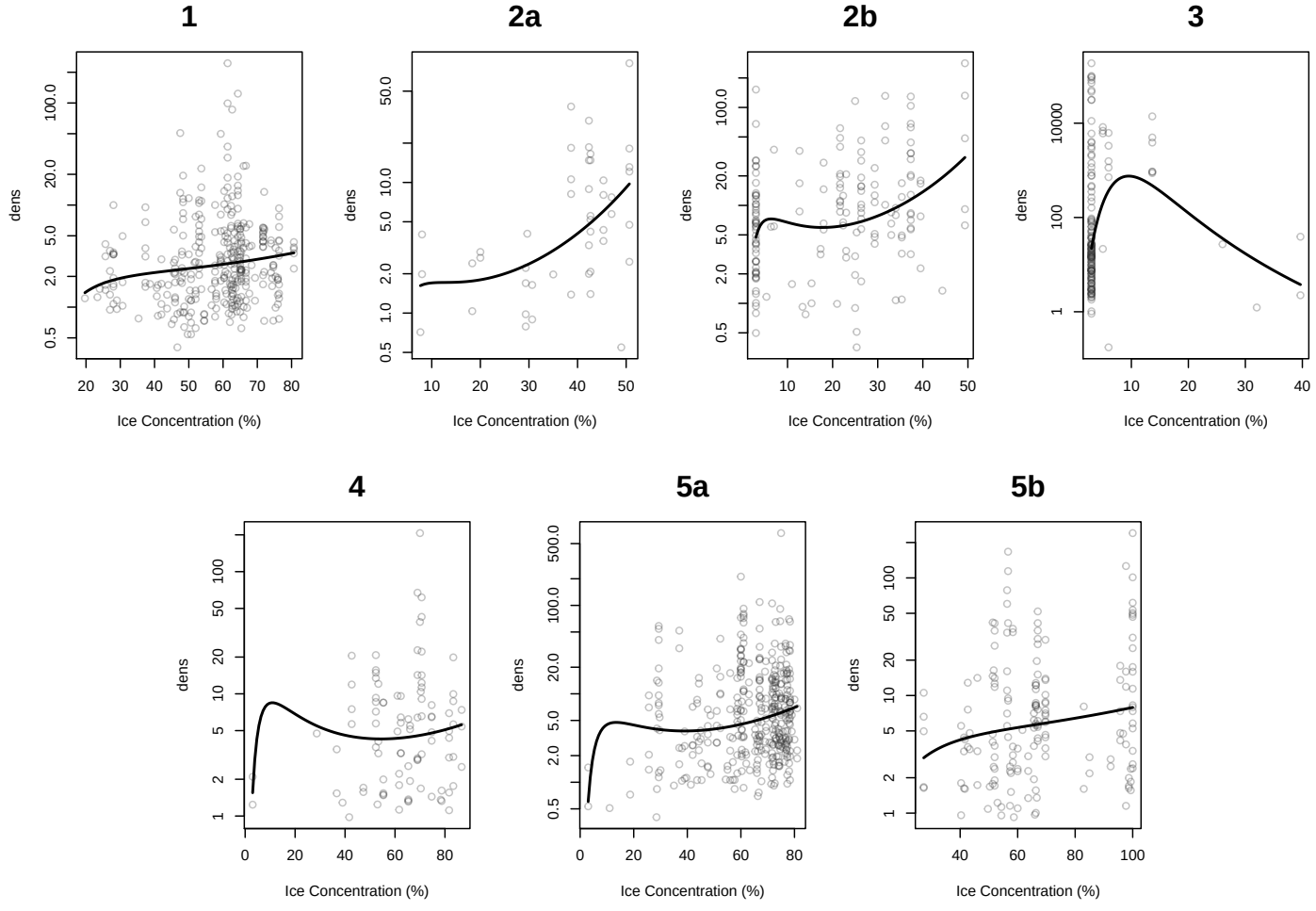
$$\texttt{lm(Dt} \sim \texttt{Shelf * (1/IC + IC + IC}^{1/2}\texttt{) * (1/DIce + DIce + DIce}^{1/2}\texttt{) - 1)}$$



Distance to ice edge



Ice Concentration



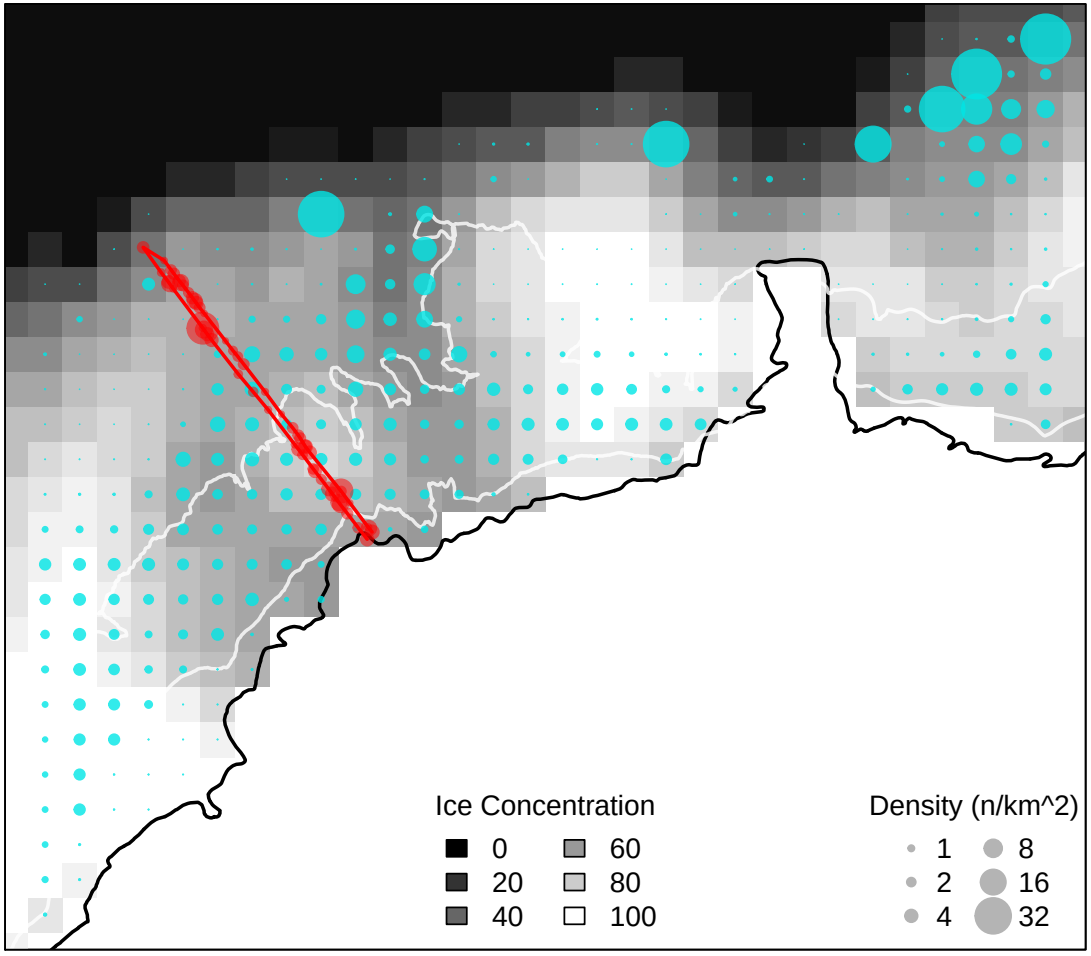
ANOVA

```
lm(Dt ~ OnShelf * (IC.r + IC + IC2) * (DIce.r + DIce + DIce2) - 1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
OnShelf	2	195.80	97.90	4898.23	0.0000	***
IC.r	1	0.42	0.42	21.18	0.0000	***
IC	1	0.05	0.05	2.31	0.1291	
IC2	1	0.03	0.03	1.33	0.2488	
DIce.r	1	0.01	0.01	0.52	0.4722	
DIce	1	0.05	0.05	2.51	0.1140	
DIce2	1	1.52	1.52	75.97	0.0000	***
OnShelf:IC.r	1	0.02	0.02	1.24	0.2672	
OnShelf:IC	1	0.00	0.00	0.00	0.9592	
OnShelf:IC2	1	0.01	0.01	0.27	0.6024	
OnShelf:DIce.r	1	0.08	0.08	4.02	0.0458	*
OnShelf:DIce	1	0.00	0.00	0.00	0.9907	
OnShelf:DIce2	1	0.01	0.01	0.57	0.4500	
IC.r:DIce.r	1	0.01	0.01	0.37	0.5409	
IC.r:DIce	1	0.01	0.01	0.52	0.4707	
IC.r:DIce2	1	0.00	0.00	0.17	0.6798	
IC:DIce.r	1	0.11	0.11	5.34	0.0214	*
IC:DIce	1	0.00	0.00	0.24	0.6251	
IC:DIce2	1	0.01	0.01	0.33	0.5665	
IC2:DIce.r	1	0.09	0.09	4.69	0.0311	*
IC2:DIce	1	0.00	0.00	0.03	0.8676	
IC2:DIce2	1	0.00	0.00	0.01	0.9202	
OnShelf:IC.r:DIce.r	1	0.05	0.05	2.74	0.0987	·
OnShelf:IC.r:DIce	1	0.05	0.05	2.34	0.1271	
OnShelf:IC.r:DIce2	1	0.18	0.18	8.79	0.0032	**
OnShelf:IC:DIce.r	1	0.02	0.02	0.92	0.3391	
OnShelf:IC:DIce	1	0.00	0.00	0.05	0.8193	
OnShelf:IC:DIce2	1	0.00	0.00	0.12	0.7286	
OnShelf:IC2:DIce.r	1	0.07	0.07	3.46	0.0635	·
OnShelf:IC2:DIce	1	0.00	0.00	0.07	0.7884	
OnShelf:IC2:DIce2	1	0.04	0.04	2.02	0.1559	
Residuals	343	6.86	0.02			

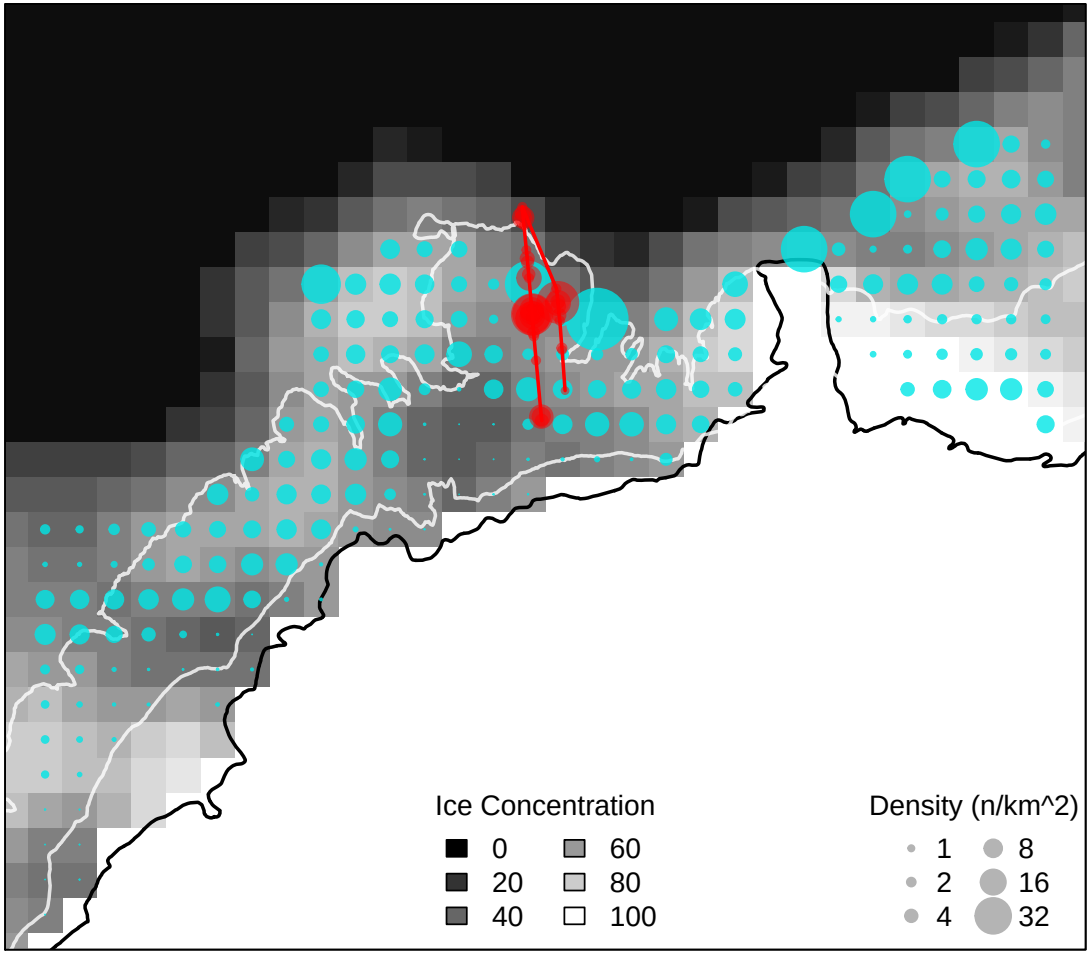
Here are some predictions:

Emage 1: 12/25/1996
estimate: 453K, CI: 181 − 3715K



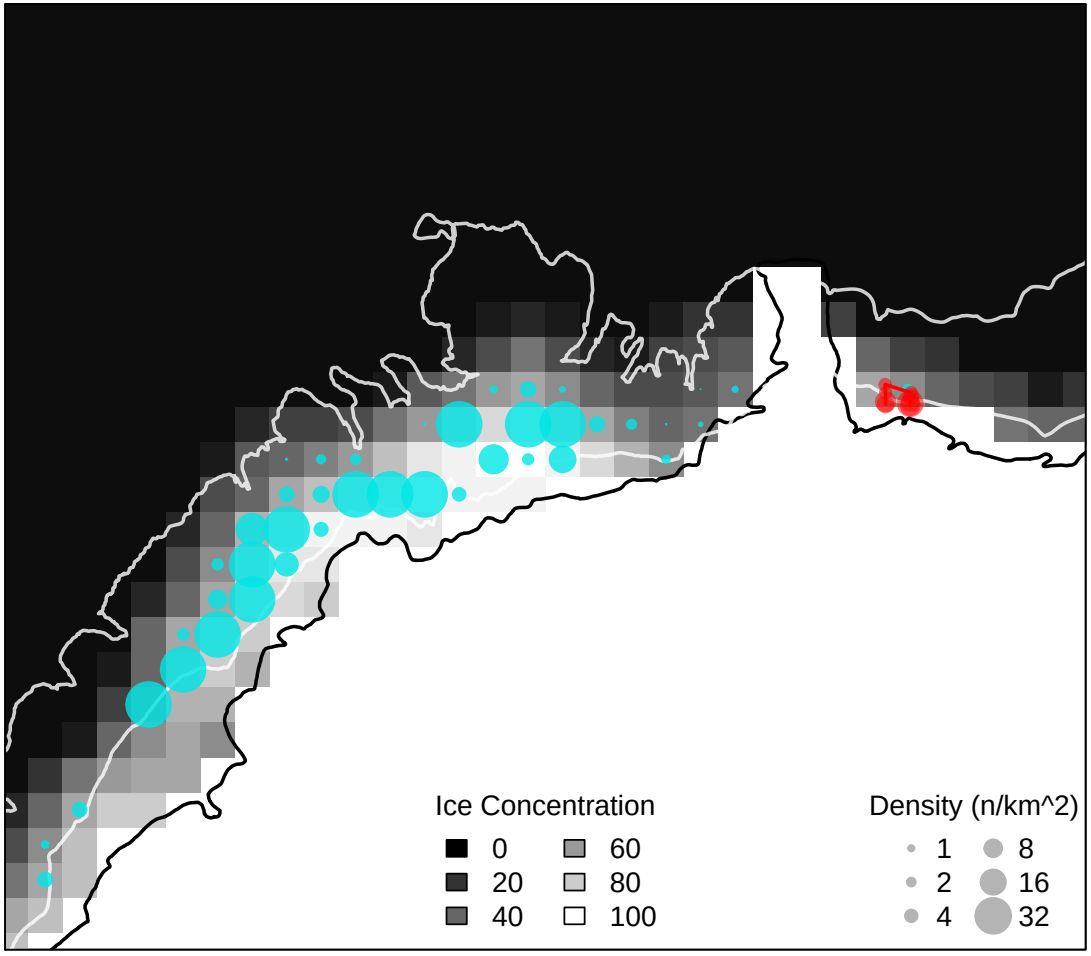
Here are some predictions:

Emage 5b: 1/11/2001
estimate: 775K, CI: 295 − 4435K

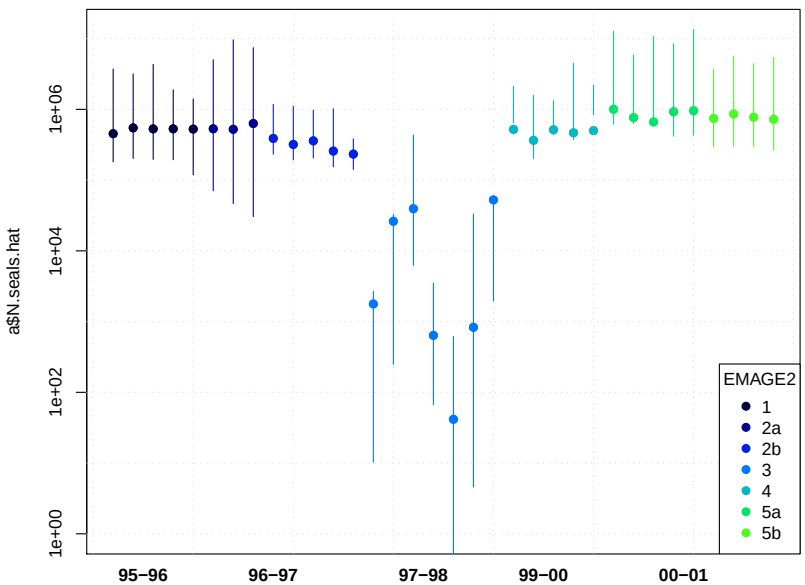


Here are some predictions:

Emage 4: 1/20/2000
estimate: 466K, CI: 374 − 4484K



$$\texttt{glm(Density}^{-0.2} \sim \texttt{Shelf (IC.r + IC + IC2) * (DIce.r + DIce + DIce2)-1}$$



	Date	N.seals	N.obs	$\widehat{N}$	C.I. (x 1000)
1	12/25/1996	58	56	453	181 - 3715
1	12/26/1996	70	63	548	202 - 3164
1	12/27/1996	94	84	530	196 - 4312
1	12/28/1996	75	69	531	194 - 1895
1	12/30/1996	75	62	527	119 - 1402
2a	1/3/1998	33	30	532	71 - 5041
2a	1/4/1998	6	6	523	47 - 9610
2a	1/5/1998	8	6	632	30 - 7452
2b	1/15/1998	19	18	388	233 - 1177
2b	1/19/1998	47	41	319	193 - 1109
2b	1/20/1998	15	14	359	205 - 979
2b	1/22/1998	39	33	258	154 - 1018
2b	1/23/1998	50	43	233	141 - 383
3	1/27/1999	4	4	2	0 - 3
3	1/28/1999	18	15	26	0 - 33
3	1/29/1999	58	43	39	6 - 434
3	1/30/1999	43	26	1	0 - 3
3	2/1/1999	14	9	0	0 - 1
3	2/7/1999	5	3	1	0 - 33
3	2/11/1999	6	4	52	2 - 54
4	1/8/2000	38	34	522	644 - 2121
4	1/10/2000	15	14	365	202 - 1581
4	1/18/2000	11	11	515	442 - 1322
4	1/20/2000	10	9	466	374 - 4484
4	1/23/2000	2	2	500	843 - 2189
5a	12/21/2000	89	81	1003	618 - 12779
5a	12/22/2000	110	101	765	629 - 5901
5a	12/24/2000	29	27	664	692 - 10764
5a	12/26/2000	63	59	925	419 - 8462
5a	12/27/2000	119	107	958	430 - 13443
5b	1/5/2001	70	65	744	293 - 3705
5b	1/9/2001	9	6	860	299 - 5652
5b	1/11/2001	36	35	775	295 - 4435
5b	1/12/2001	63	52	725	264 - 5460