# Linear Regression

Eli Gurarie

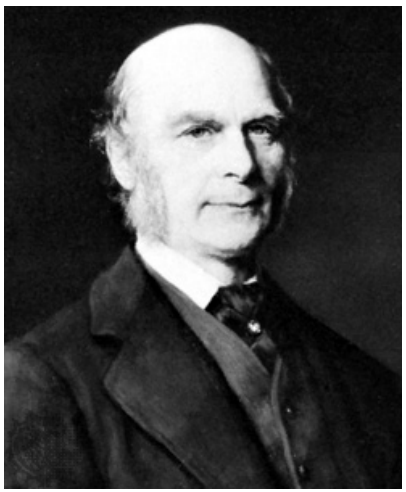
QERM 598 - Lecture 4

University of Washington - Seattle

January 29, 2008

# Historical roots of Linear Regression

Linear regression owes much to Sir Francis Galton (1822  1911), a half-cousin of Charles Darwin and one of a generation of basically brilliant English Victorian polymaths. He made important contributions to anthropology, geography, meteorology, genetics, psychometrics and statistics.



Galton was really, really into counting and quantifying things, inventing the *questionnaire* along the way to assess intelligence. He noted that 'exceptional' parents produce more 'mediocre' children (and, interestingly, vice versa!). Hence the idea of 'regression' (as in regression to mediocrity). This slightly misleading name has stuck to a very useful statistical tool to this day.

His contributions were truly many and diverse (note: the dog whistle! forensic fingerprinting! the Galton-Watson stochastic process!) Fortunately, some of his greatest passions, *eugenics* and *phrenology*, never got too far off the ground.

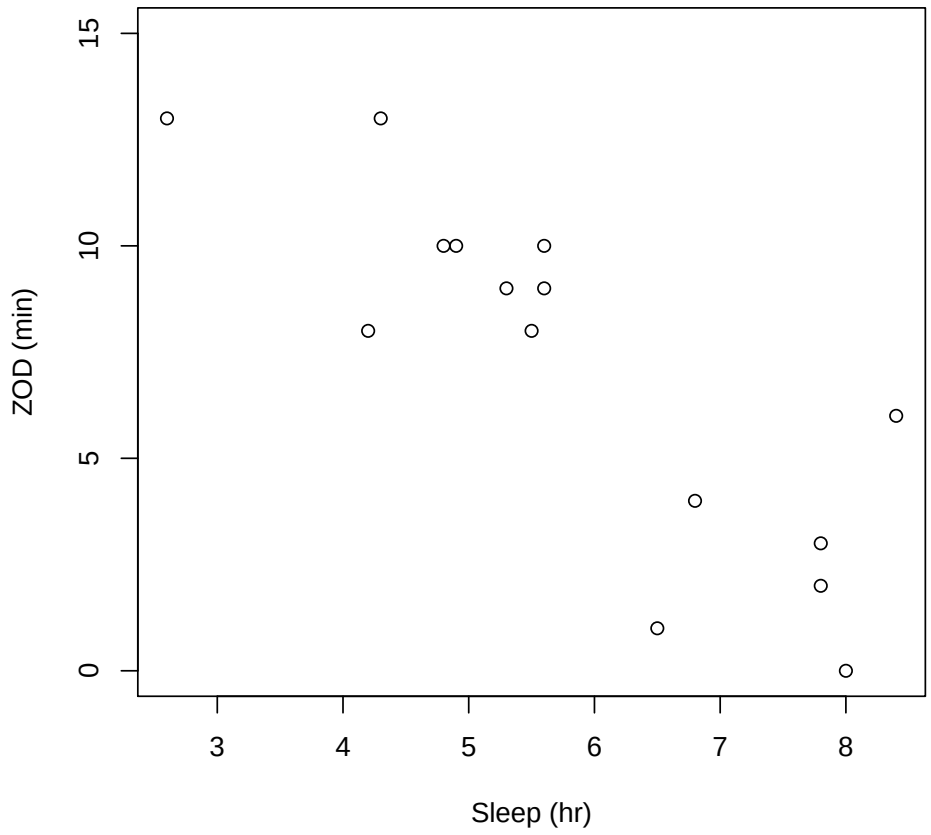
# The Great QERM Insomnia Experiment

One Wednesday during a 50 min Quantitaive Fisheries talk given by a fellow student, the zone-outs duration (ZOD) of fifteen other students was carefully recorded by an investigator. Later, the investigator collected data on the quantity of sleep the students received the preceding evening.
The results (in minutes) 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

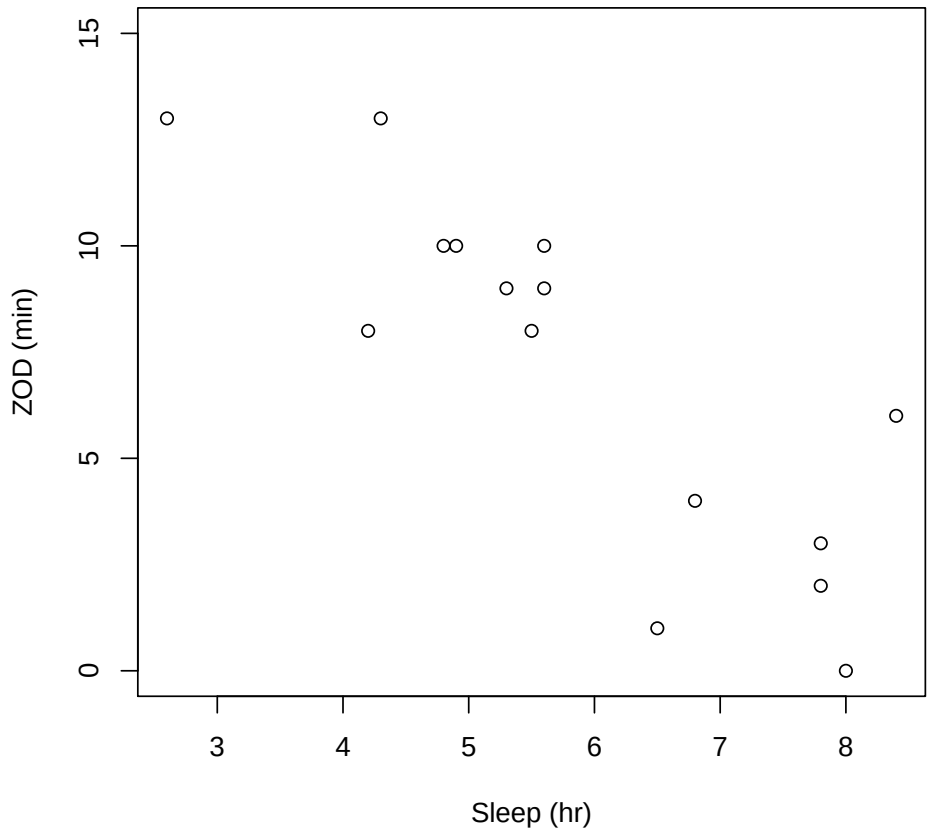
Looks like there might be a relationship!




Perhaps linear?

# Results

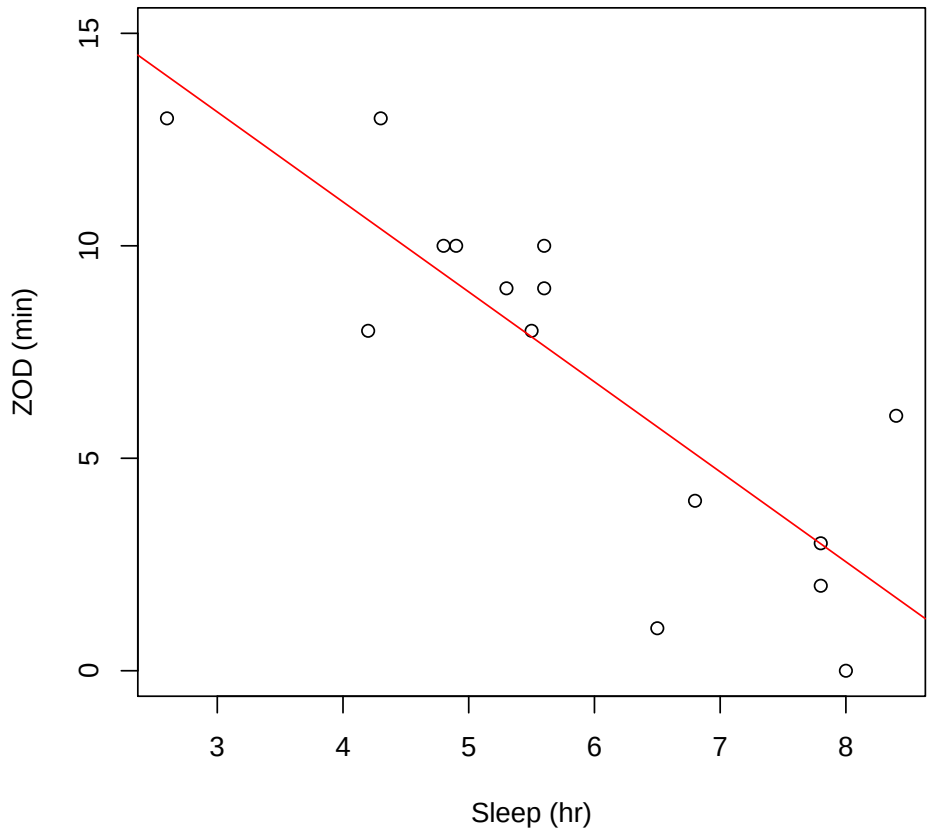
Looks like there might be a relationship!




Perhaps linear?

# Results

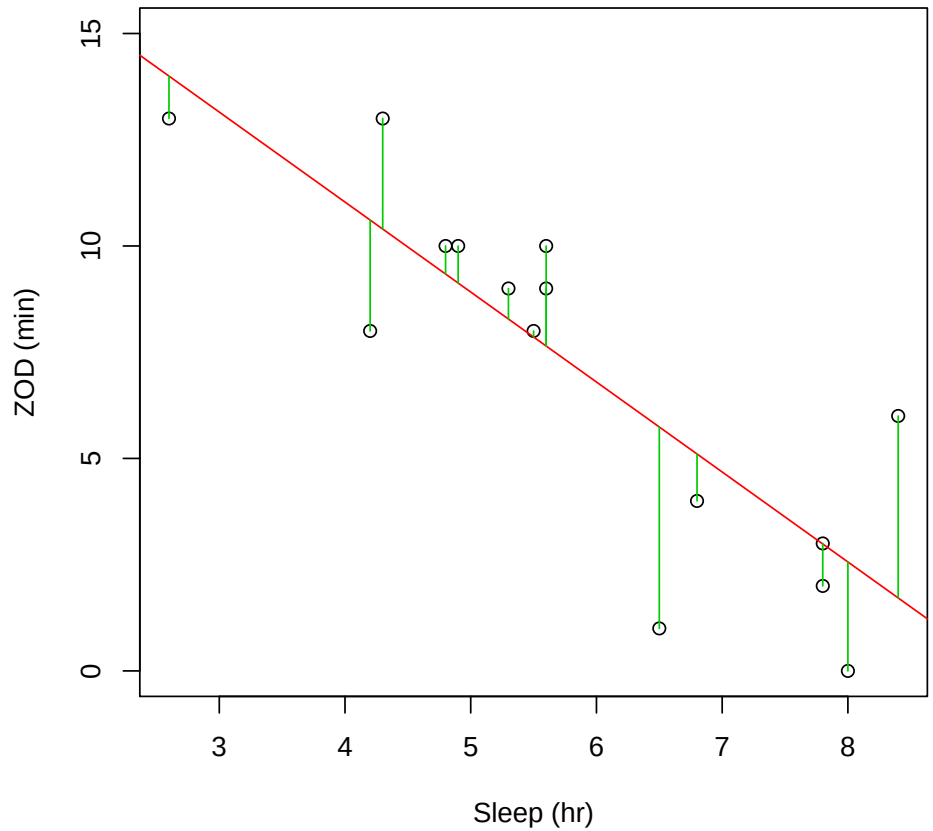
We propose a simple linear model: $Y_i = \alpha X_i + \beta + \epsilon_i$




If $\epsilon_i \sim$ Normal $\left\{0, \sigma^2\right\}$ AND ARE iid we can apply the simplest form of *linear regression* .

# Method of least squares

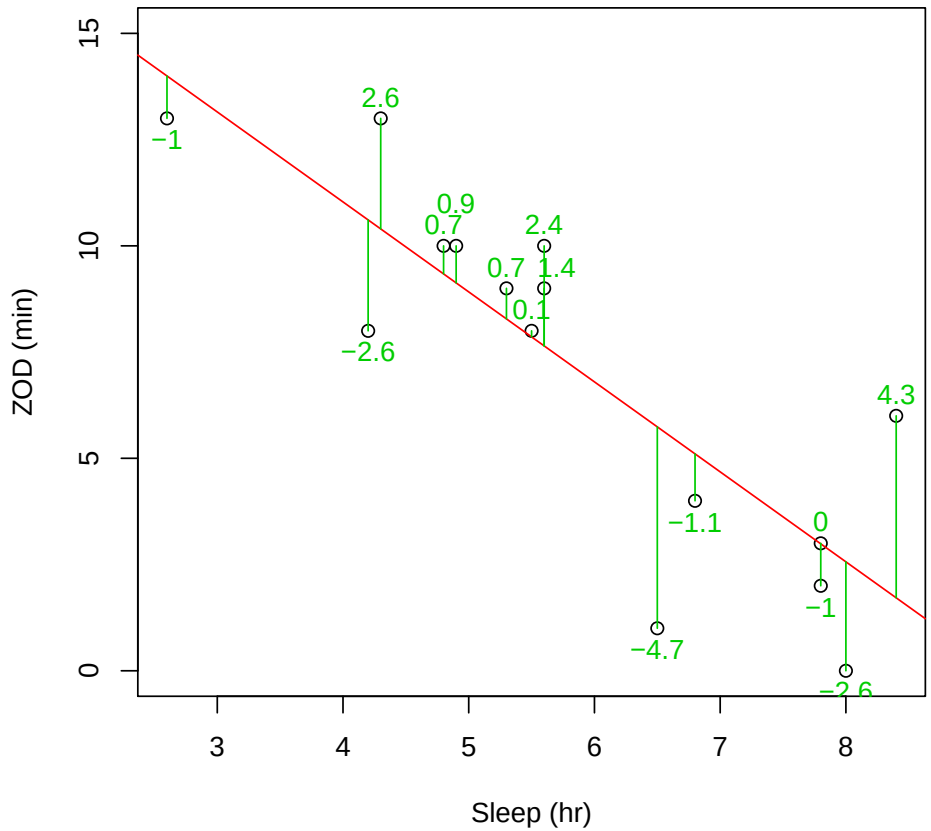
Draw some line (model)!




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

# Method of least squares

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




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

# How?

With a little math! Minimization means:

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

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

Solving these two equations yields:

$$\hat{\alpha} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \quad (3)$$

$$\hat{\beta} = \bar{Y} - \hat{\alpha}\bar{X} \quad (4)$$

# How?

With a little math! Minimization means:

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

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

Solving these two equations yields:

$$\hat{\alpha} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \quad (3)$$

$$\hat{\beta} = \bar{Y} - \hat{\alpha}\bar{X} \quad (4)$$

The estimates:

$$\hat{\alpha} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \quad (5)$$

$$\hat{\beta} = \bar{Y} - \hat{\alpha}\bar{X} \quad (6)$$

Plug in our numbers:

$$\begin{aligned} \bar{X} &= 5.87 \quad (7) \\ \bar{Y} &= 7.07 \\ \hat{\alpha} &= -81/38.2 = -2.11 \\ \hat{\beta} &= 7.07 + 2.11 * 5.87 = 19.5 \end{aligned}$$

Our minimized SSE is:

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

Our final model is

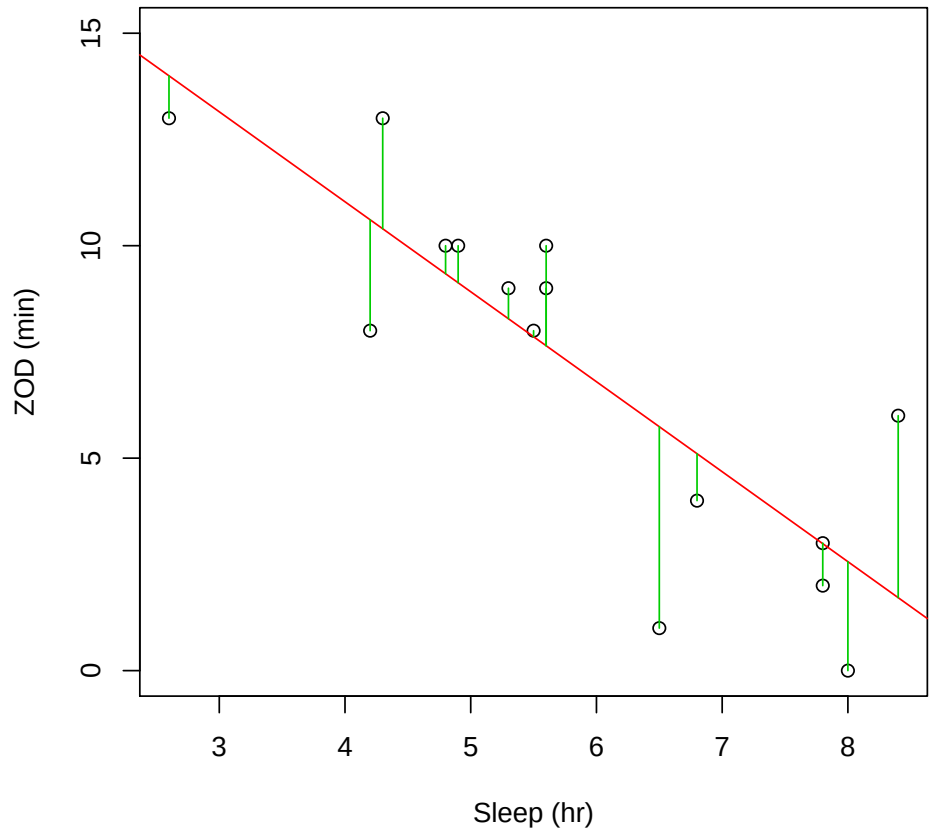
$$Y_i = 19.5 - 2.11X_i + \epsilon \quad (8)$$

with

$$\hat{\sigma}^2 = MSE = SSE/(n-2) = 5.64 \quad (9)$$

# And

Of course, it fits quite nicely:




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

# Inference

How is $\hat{\alpha}$ distributed? Rewrite $\alpha$ as:

$$\hat{\alpha} = \sum k_i Y_i \text{; Where: } k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \tag{10}$$

(note: $\sum k_i = 0$ , $\sum k_i X_i = 1$ and $\sum k_i^2 = 1/\sum (X_i - \bar{X})^2$ )

In our model, $Y$ are iid and normally distributed... then $\sum k_i Y_i$ is also normally distributed. It's expected value is:

$$\mathrm{E}[\hat{\alpha}] = \mathrm{E}\left[\sum k_i Y_i\right] = \sum k_i(\alpha X_i + \beta) = \beta \sum k_i + \alpha \sum k_i X_i = \alpha \tag{11}$$

Therefore: $\hat{\alpha}$ is unbiased!

# Inference II

What about the variance of $\alpha$?

$$\hat{\alpha} = \sum k_i Y_i; \text{ Where: } k_i = \frac{X_i - \bar{X}}{\sum X_i - \bar{X}^2} \tag{12}$$

(note: recall that if $Z_1, Z_2, ..., Z_n$ are iid $\mathrm{N}\left\{\mu, \sigma^2\right\}$ and $a_i$ are constants:)

$$\sum_{i=1}^{n} a_i Z_i \sim \mathrm{N}\left\{\sum a_i, \sum a_i^2 \sigma^2\right\} \tag{13}$$

Recalling that $\mathrm{Var}\left[Y_i\right] = \sigma^2$ , the variance of $\hat{\alpha}$ (defined as $\sigma^2[\alpha]$ ):

$$\sigma^2[\alpha] = \sigma^2 \sum k_i^2 = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$$

And the estimate of the $\mathrm{Var}\left[\alpha\right]$ (defined as $s^2[a]$ ) is:

$$s^2[a] = \hat{\sigma}^2 \sum k_i^2 = \frac{MSE}{\sum (X_i - \bar{X})^2} \tag{14}$$

# The Test Statistic

Now that we have normal variable with an estimate for the mean and the variance, we can perform the following test:

$$\mathrm{H}_0 : \hat{\alpha} = 0$$
$$\mathrm{H}_1 : \hat{\alpha} \neq 0$$

Using the test statistic:

$$t_0 = \frac{\hat{\alpha}}{s(\hat{\alpha})}$$

How is this statistic distributed?:

$$\begin{aligned} \frac{\hat{\alpha} - \alpha}{\sigma[\alpha]} &\sim \mathrm{N}\{0, 1\} \\ SSE/(n-2)\sigma^2 &\sim \chi^2(n-2) \\ \frac{s^2[\alpha]}{\sigma^2[\alpha]} = \frac{MSE/\sum(X_i - \bar{X})^2}{\sigma^2/\sum(X_i - \bar{X})^2} &= MSE/\sigma^2 \sim \chi^2(n-2) \end{aligned}$$

Thus, under the null hypothesis, $t_0$ is the ratio of a Normal and Chi-squared, i.e.:

$$t_0 \sim \mathrm{T}(n-2) \tag{15}$$

# Perform the test

Our test statistic:

$$t_0 = \frac{\hat{\alpha}}{s(\hat{\alpha})} = \frac{-2.12}{\sqrt{0.147}} = -5.5$$

Compare to a $T(n-2)$ ... p-value $= 9.89e-05$ . And so, we:

- (a) REJECT THE NULL HYPOTHESIS ;
- (b) conclude that the SLOPE is NOT equal to 0 ;
- (c) conclude that there IS an effect of Sleep amount on Zone-Out Duration ;
- (d) have collected a small but convincing piece of evidence for some grand theory of QERM student concentration ability.

# Some comments

- *Linear* regression is NOT necessarily a straight line response to a factor. Linearity refers to the fact that the response variable is modeled as a linear (i.e. additive) combination of parameters and factor responses. Thus: $Y_i = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ is a quadratic response of $Y$, but the model is linear and all the linear regression theory applies, as long as $\epsilon$ is iid, normal, etc.
- The estimates which we obtained using straightforward algebra and calculus can also be obtained using maximum likelihood theory, which is in some ways stronger in it's conclusions.
- We brushed over an assessment of the assumptions, but as with ANOVA it is important to look at these! Also, we didn't discuss inference on the intercept parameter $\beta$. However, this is generally a parameter of lesser interest and it's distribution is straightforward to derive from the distribution of $\alpha$, which is why it's left as a homework exercise!

# Some comments

- *Linear* regression is NOT necessarily a straight line response to a factor. Linearity refers to the fact that the response variable is modeled as a linear (i.e. additive) combination of parameters and factor responses. Thus: $Y_i = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ is a quadratic response of $Y$, but the model is linear and all the linear regression theory applies, as long as $\epsilon$ is iid, normal, etc.
- The estimates which we obtained using straightforward algebra and calculus can also be obtained using maximum likelihood theory, which is in some ways stronger in it's conclusions.
- We brushed over an assessment of the assumptions, but as with ANOVA it is important to look at these! Also, we didn't discuss inference on the intercept parameter $\beta$. However, this is generally a parameter of lesser interest and it's distribution is straightforward to derive from the distribution of $\alpha$, which is why it's left as a homework exercise!

# Some comments

- *Linear* regression is NOT necessarily a straight line response to a factor. Linearity refers to the fact that the response variable is modeled as a linear (i.e. additive) combination of parameters and factor responses. Thus: $Y_i = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ is a quadratic response of $Y$, but the model is linear and all the linear regression theory applies, as long as $\epsilon$ is iid, normal, etc.
- The estimates which we obtained using straightforward algebra and calculus can also be obtained using maximum likelihood theory, which is in some ways stronger in it's conclusions.
- We brushed over an assessment of the assumptions, but as with ANOVA it is important to look at these! Also, we didn't discuss inference on the intercept parameter $\beta$. However, this is generally a parameter of lesser interest and it's distribution is straightforward to derive from the distribution of $\alpha$, which is why it's left as a homework exercise!

# Some comments

- *Linear* regression is NOT necessarily a straight line response to a factor. Linearity refers to the fact that the response variable is modeled as a linear (i.e. additive) combination of parameters and factor responses. Thus: $Y_i = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ is a quadratic response of $Y$, but the model is linear and all the linear regression theory applies, as long as $\epsilon$ is iid, normal, etc.
- The estimates which we obtained using straightforward algebra and calculus can also be obtained using maximum likelihood theory, which is in some ways stronger in it's conclusions.
- We brushed over an assessment of the assumptions, but as with ANOVA it is important to look at these! Also, we didn't discuss inference on the intercept parameter $\beta$. However, this is generally a parameter of lesser interest and it's distribution is straightforward to derive from the distribution of $\alpha$, which is why it's left as a homework exercise!

# Some comments II

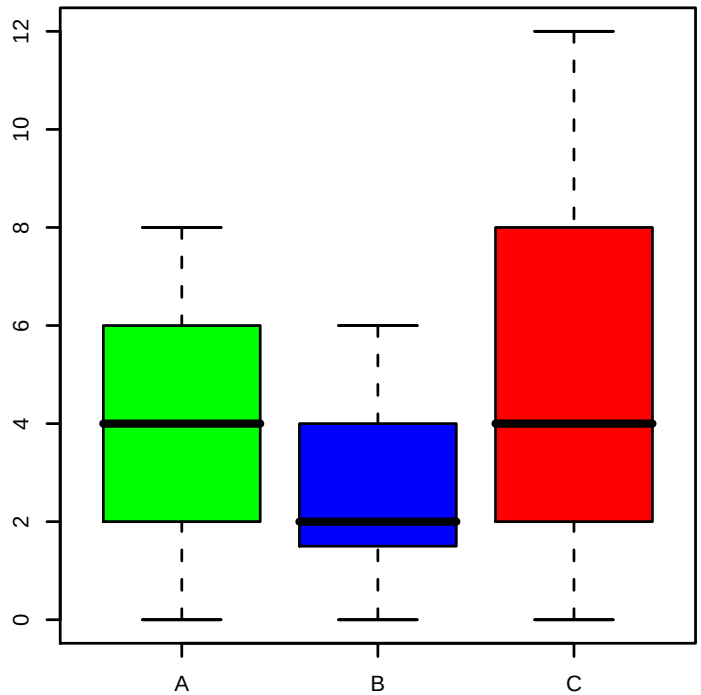
- The basic form of the linear regression model looks somewhat like the single-factor ANOVA we looked at last week. This is not a coincidence! Both are special cases of a greater regression theory and their relationship becomes clear when the models are rewritten in matrix notation ( $Y = \mathbf{M}$ X ). That's all I'll say now. For details, I defer to QERM 514 and the STAT 570's series.
- However, I would like to expose you to a more complicated model that combines linear regression and single factor ANOVA ...

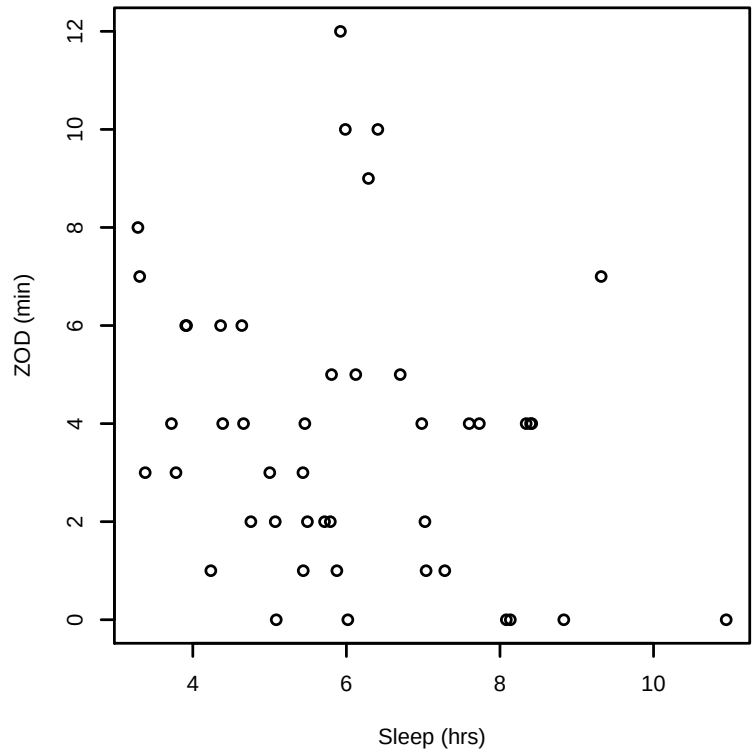
# Some comments II

- The basic form of the linear regression model looks somewhat like the single-factor ANOVA we looked at last week. This is not a coincidence! Both are special cases of a greater regression theory and their relationship becomes clear when the models are rewritten in matrix notation ( $Y = \mathbf{M}\ X$ ). That's all I'll say now. For details, I defer to QERM 514 and the STAT 570's series.
- However, I would like to expose you to a more complicated model that combines linear regression and single factor ANOVA ...

# The Great QERM Insomnia Pie Interaction Experiment

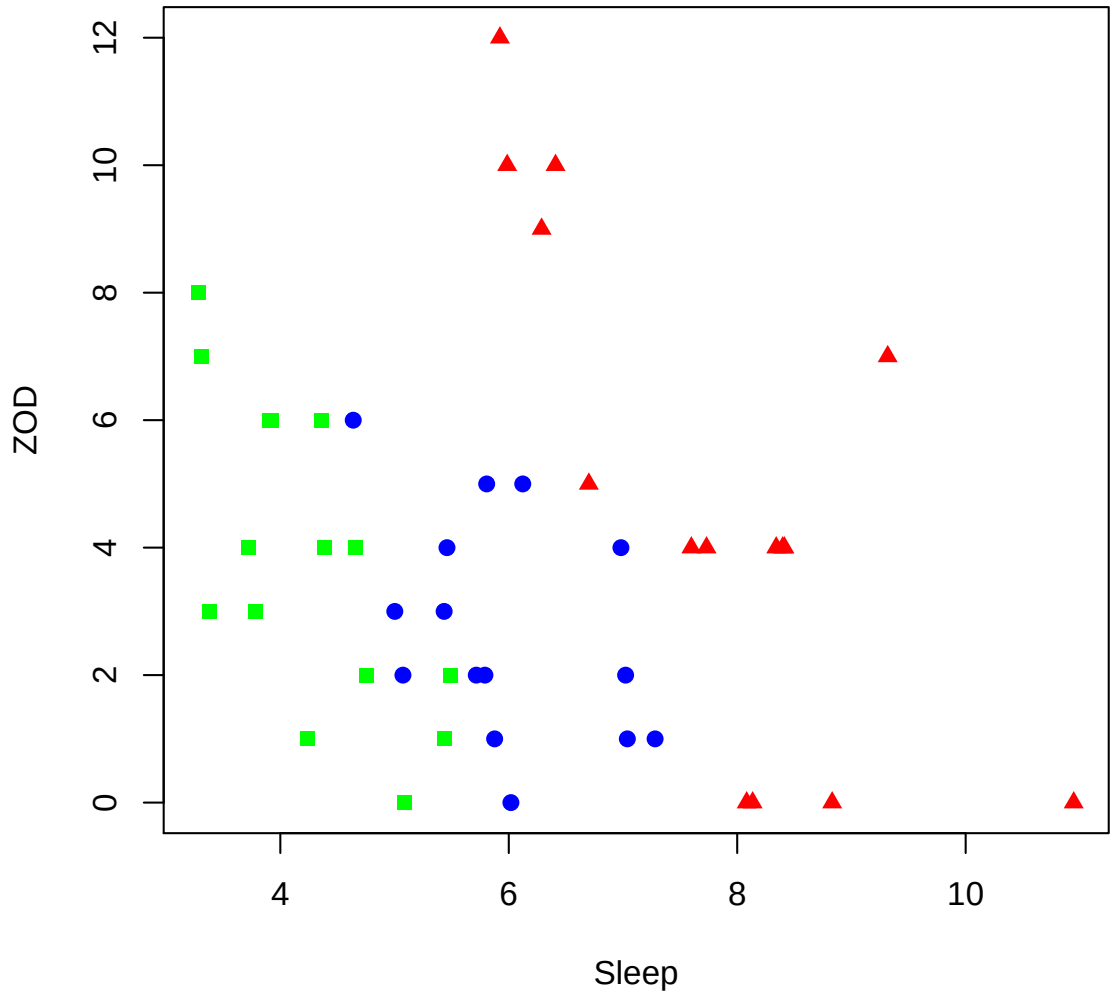
In order to integrate the conclusions of the previous two experiments, a researcher has decided to perform the following experiment: over three week, each of fifteen QERM students are randomly assigned one of three pies (apple , blueberry and cherry ) to eat and the subsequent zone-out duration during a linear-regression class is monitored. The results are plotted below:




What are the patterns? Tough to tell!

A more informative plot:



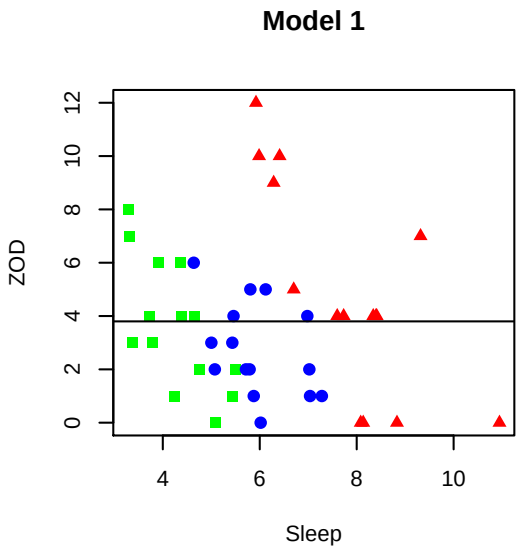
# Models

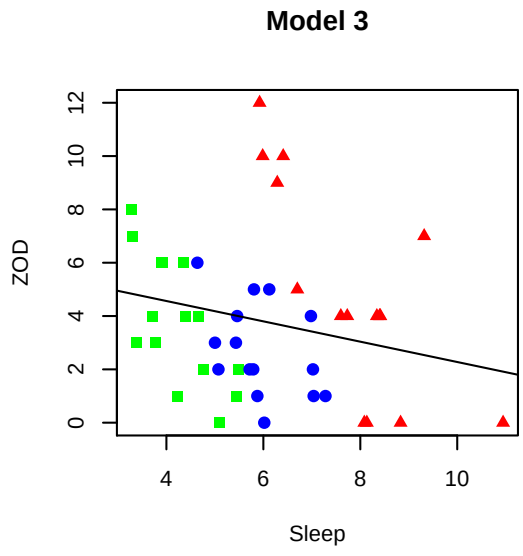
Consider the following models:

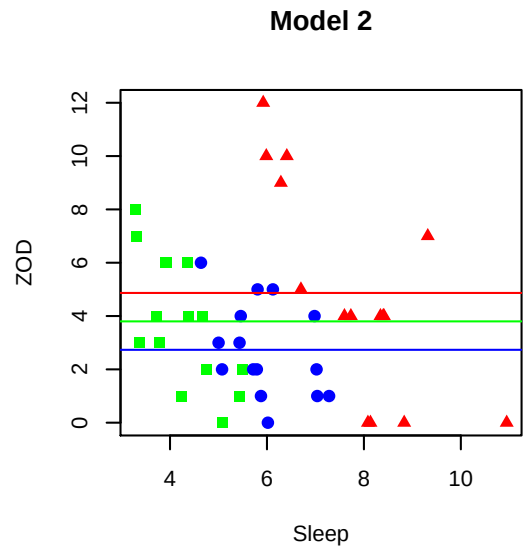
$$
\begin{aligned}
&1: \; Y_i = \alpha + \epsilon_i && \text{where } i \in \{1, 2, ...N\} \\
&2: \; Y_i = \alpha + \beta X_i + \epsilon_i && \\
&3: \; Y_{ij} = \alpha + \gamma_i + \epsilon_{ij} && \text{where } i \in \{1, 2, ...a\} \\
&4: \; Y_{ij} = \alpha_i + \beta X_{ij} + \gamma_i + \epsilon_{ij}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

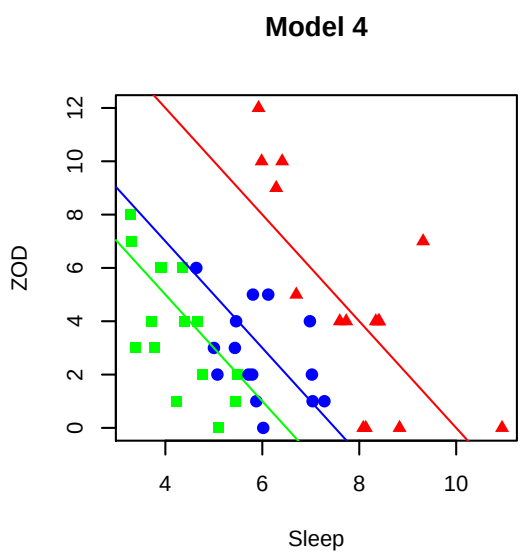
where $N$ is the total number of observations, $a$ is the number of groups (Pies), $n$ is the number of measurements per group, $\alpha$ is an intercept parameter, and $\beta$ is a slope parameter, and $\gamma$ is a group effect.

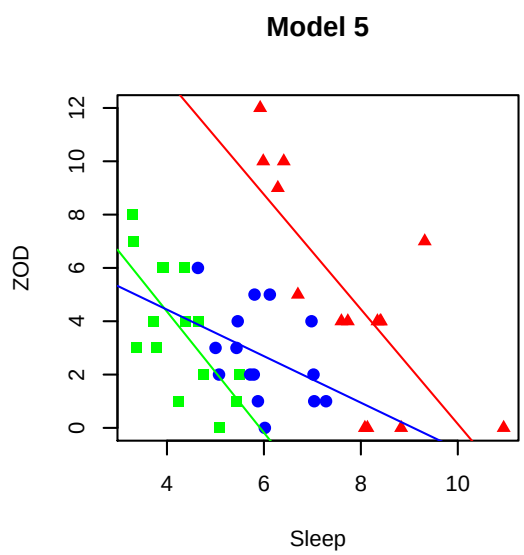
# Models


Model 1
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 2
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 (obtained in one line of R code):

| | 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 it's parameter values? How do we interpret and report the results?