Models, parameters, and the General Linear Model

Notation used in your book

Population parameters	Usual null hypotheses	Sample estimates
μ, σ^2	$\mu = 0$	$\bar{y}, s^2$
$\mu_A, \mu_B, \mu_C, \sigma^2$	$\mu_A = \mu_B = \mu_C$	$\bar{y}_A, \bar{y}_B, \bar{y}_C, s^2$
α, β, σ^2	$\beta = 0$	a, b, s^2

Parameters: True population values that are unknown, but are estimated from sample data – denoted with Greek alphabetic symbols

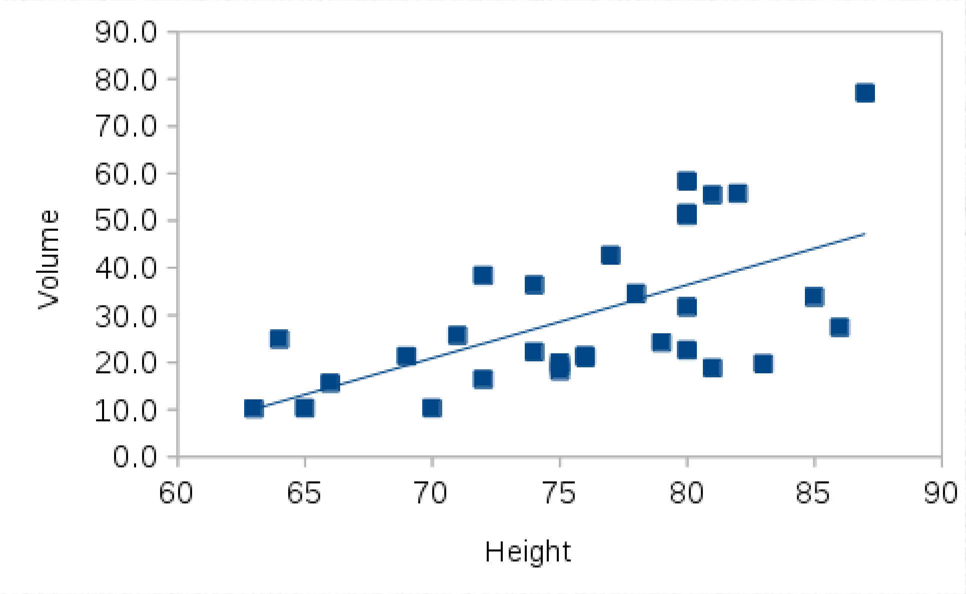
Estimates: Statistics calculated from data that estimate population parameters – denoted in Latin alphabetic symbols

Fitting models to data

- All of the statistical procedures we have learned (and will learn in here) are model-based
- In each case, we can write an equation relating a mean response to values of one or more predictor variables
- A model-based analysis of data is done by estimating and intepreting model parameters = model coefficients (intercepts and slopes)

The obvious case: regression

$$\hat{y} = \alpha + \beta x$$



Estimates of coefficients:

a = -87.12

b = 1.54

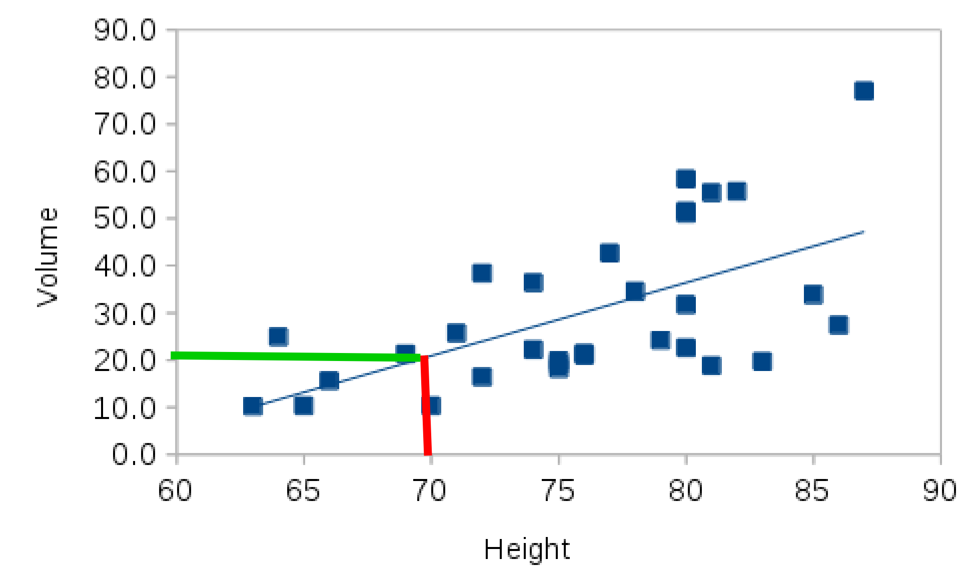
Regression equation:

Volume = -87.12 + 1.54 Height

If the line explains enough variation in volume to be statistically significant, we interpret the results by interpreting the slope – which means...what?

Predicted values on a regression line

- Predicted values are means for y at a given value on the x-axis
- Predicted by plugging x into regression equation
- To predict the volume of lumber in a 70 ft tree:



Regression equation:

Volume = -87.12 + 1.54 Height

Volume = -87.12 + 1.54 (70) = 20.68 ft^3

GLM – ANOVA is also regression

- ANOVA can be expressed as a special case of the linear regression model

Regression	***ANOVA***
Response = Predictor *Both variables are numeric*	*Response = Predictor* *Numeric response, categorical predictor*

- In both cases we ask, "Does the mean of the response variable depend on the value of the predictor variable?"
- Both ANOVA and regression are thus special cases of the General Linear Model (GLM)

Typical ANOVA data set

- Chick weights (response, g) fed on one of six different feeds (predictor)
- How do we make the predictor numeric so we can use regression to analyze the data?

weight	feed
179	horsebean
160	horsebean
…	…
309	linseed
229	linseed
…	…
243	soybean
230	soybean
…	…
423	sunflower
340	sunflower
…	…
325	meatmeal
257	meatmeal
…	…
368	casein
390	casein
…	…

How about this?

- Assign a number to each level
- Use the numbers as the predictor in a regression
- R will let you do this, but does it give you the same results as an ANOVA of the data?

weight	feed	feed.num
179	horsebean	1
160	horsebean	1
…	…	...
309	linseed	2
229	linseed	2
…	…	...
243	soybean	3
230	soybean	3
…	…	...
423	sunflower	4
340	sunflower	4
…	…	...
325	meatmeal	5
257	meatmeal	5
…	…	...
368	casein	6
390	casein	6
…	…	...

No, not that

Why not? Assigning numeric codes to feed types does not make feed a numeric variable

And, why alphabetical order, anyway?

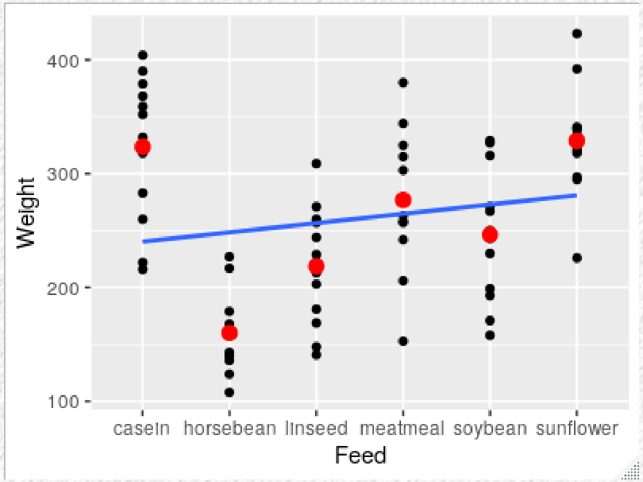
Feed is a nominal categorical variable - any order for Feed is valid, but would give different results for the regression

ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
feed	5	231129	46226	15.365	5.936e-10
Residuals	65	195556	3009		

Regression

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
feed	1	13893	13892.5	2.3222	0.1321
Residuals	69	412793	5982.5		



ANOVA as a regression the right way

- Converting categories to numeric predictors is called **coding**
- Dummy coding (a.k.a. **treatment contrasts**) is what R uses by default
- We'll start with just two of the bean types and dummy code them

weight	feed	horsebean
179	horsebean	1
160	horsebean	1
136	horsebean	1
227	horsebean	1
217	horsebean	1
168	horsebean	1
108	horsebean	1
124	horsebean	1
143	horsebean	1
140	horsebean	1
309	linseed	0
229	linseed	0
181	linseed	0
141	linseed	0
260	linseed	0
203	linseed	0
148	linseed	0
---	---	---

Dummy coding feed

Create a new *numeric* variable named for one of the **levels** of feed (called “horsebean” here)

Record a 1 in horsebean column when feed type is horsebean, 0 when it is not (i.e. when it is linseed)

Run a regression with **horsebean** as the predictor variable, **weight** as the response variable

$$weight = horsebean$$

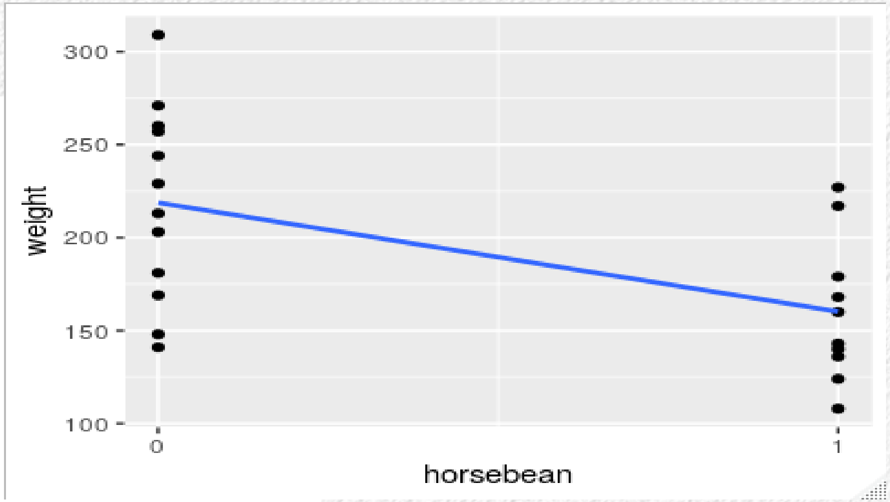
Model formula

$$\widehat{weight} = \alpha + \beta \, horsebean$$

Regression equation

Feed types analyzed as a regression

Regression Analysis:
Weight versus horsebean




```
The regression equation is

weight = 218.75 – 58.55 horsebean


Response: weight

          Df Sum Sq Mean Sq F value   Pr(>F)

horsebean  1  18699 18698.7  8.6086 0.008205

Residuals 20  43442  2172.1
```

If this is correct, it will be identical to an ANOVA – is it?

ANOVA of feed same as regression of horsebean

```
As an ANOVA, feed type as categorical predictor
Response: weight
          Df Sum Sq Mean Sq F value   Pr(>F)
feed       1  18699 18698.7  8.6086 0.008205
Residuals 20  43442  2172.1
```

```
As a regression, horsebean as numeric predictor
Response: weight
          Df Sum Sq Mean Sq F value   Pr(>F)
horsebean  1  18699 18698.7  8.6086 0.008205
Residuals 20  43442  2172.1
```

They match!

Only the name of the predictor is different

Coefficients are not group means, predicted values are

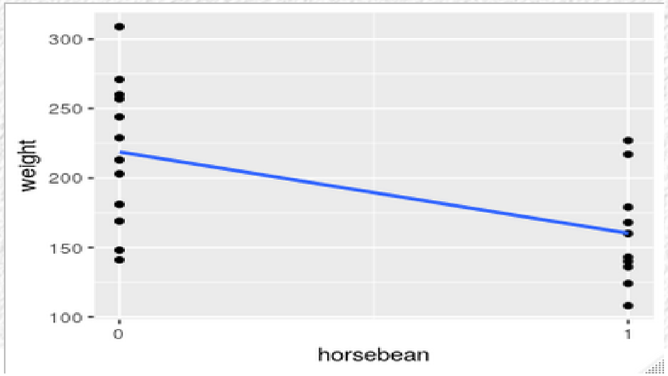
- The coefficients from the regression are:

  Intercept = 218.75, slope = -58.55

- The mean weight for the feed groups are:

  horsebean: 160.2, linseed = 218.75

- The intercept is the mean for linseed, but the slope is not the mean for horsebean
- We need to predict the mean for horsebean from the regression equation to get its mean



Predicted values as group means

Regression Analysis: weight = horsebean

```
The regression equation is
weight = 218.75 - 58.55 horsebean
```

$$\hat{y}_0 = 218.75 - 58.55(0) = 218.75$$
$$\hat{y}_1 = 218.75 - 58.55(1) = 160.20$$

Mean weights by feed

feed	mean
horsebean	160.20
linseed	218.75

Predicted values are mean of y (weight) at a given x (horsebean)

Intercept coefficient is the horsebean = 0 mean (linseed)

Slope coefficient is the difference between linseed and horsebean

Therefore, regression coefficients are not all group means, but predicted values are

R provides tests of coefficients

t value is coeff estimate divided by se

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   218.75      13.45  16.259 5.39e-13
horsebean     -58.55      19.96  -2.934  0.00821
```

Tests if linseed mean is different from 0 (not interesting)

Tests difference between horsebean and linseed (what we want to know!)

Extending this approach to 6 feed types

- Can use the same approach with more than 2 groups, but need additional dummy-coded columns to do it
- With 6 feeds we need 5 columns (in general, one fewer than the number of levels)
  - One feed is set as baseline group (first alphabetically by default)
  - One column created for each of the other 5 feeds
    - Enter a 1 for a row if the feed level matches the column name
    - Enter a 0 otherwise
  - A 0 across all five columns is used for the baseline group
- These five columns are then used as predictors in a **multiple regression**

weight	feed	horsebean	linseed	meatmeal	soybean	sunflower
368	casein	0	0	0	0	0
390	casein	0	0	0	0	0
…	…	…	…	…	…	…
179	horsebean	1	0	0	0	0
160	horsebean	1	0	0	0	0
…	…	…	…	…	…	…
309	linseed	0	1	0	0	0
229	linseed	0	1	0	0	0
…	…	…	…	…	…	…
325	meatmeal	0	0	1	0	0
257	meatmeal	0	0	1	0	0
…	…	…	…	…	…	…
243	soybean	0	0	0	1	0
230	soybean	0	0	0	1	0
…	…	…	…	…	…	…
423	sunflower	0	0	0	0	1
340	sunflower	0	0	0	0	1
…	…	…	…	…	…	…

Which feed is the baseline?

Does each feed level only get a 1 in its matching column?

Now we use the dummy coded columns in a multiple regression model

Multiple regression

- Multiple regression extends the simple linear regression model
- Still a single intercept, and each predictor is still multiplied by a slope, but these products are added across predictors

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_k x_k$$

- Each dummy variable will be used as a predictor – we will have a single intercept, but will have one slope for each dummy variable

Model formula: weight = horsebean + linseed + meatmeal + soybean + sunflower

feed	horsebean	linseed	meatmeal	soybean	sunflower
casein	0	0	0	0	0
horsebean	1	0	0	0	0
linseed	0	1	0	0	0
meatmeal	0	0	1	0	0
soybean	0	0	0	1	0
sunflower	0	0	0	0	1

```
Coefficients:
             Estimate
(Intercept)  323.583
horsebean   -163.383
linseed     -104.833
meatmeal     -46.674
soybean      -77.155
sunflower      5.333
```

Multiple regression equation:

$$Intercept + \beta_{horsebean}\, horsebean + \beta_{linseed}\, linseed + \beta_{meatmeal}\, meatmeal + \beta_{soybean}\, soybean + \beta_{sunflower}\, sunflower = \hat{weight}$$

```
casein     = 323.6 – 163.4 (0) – 104.8 (0) – 46.7 (0) – 77.2 (0) + 5.3 (0) = 323.6
horsebean  = 323.6 – 163.4 (1) – 104.8 (0) – 46.7 (0) – 77.2 (0) + 5.3 (0) = 160.2
linseed    = 323.6 – 163.4 (0) – 104.8 (1) – 46.7 (0) – 77.2 (0) + 5.3 (0) = 218.8
meatmeal   = 323.6 – 163.4 (0) – 104.8 (0) – 46.7 (1) – 77.2 (0) + 5.3 (0) = 276.9
soybean    = 323.6 – 163.4 (0) – 104.8 (0) – 46.7 (0) – 77.2 (1) + 5.3 (0) = 246.4
sunflower  = 323.6 – 163.4 (0) – 104.8 (0) – 46.7 (0) – 77.2 (0) + 5.3 (1) = 328.9
```

Each predicted value is a group mean

Intercept is casein mean

Slopes are differences between casein and that feed

The General Linear Model (GLM)

- The General Linear Model is thus:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

- General because it encompasses:
  - ANOVA
  - Regression
  - Multiple predictors, including mixes of categorical and numeric ones
- Linear because it uses a series of variables multiplied by coefficients, added together

Problem: we don’t actually want our factors to be regression predictors

- We can run an ANOVA as a multiple regression, but we still want to compare group means
- How do we recover comparisons between means from a model with dummy-coded predictors?
- Solution: construct the ANOVA table by adding SS and d.f. across the dummy coded predictors
- GLM is used to get the ANOVA table, can still follow up with post-hocs to find out which means differ

Building an ANOVA table for the categorical predictor from a GLM

- Each predictor is given a regression-style row in the ANOVA table for multiple regression
  - Each has SS explained by the predictor
  - Each has 1 d.f.
- To convert this to a factor MS for the feed variable:
  - Add SS across predictors
  - Add df across predictors
- This is usually done automatically for you by stats packages

Response: weight

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
horsebean	1	118991	118991	39.5510	3.064e-08	***
linseed	1	52241	52241	17.3640	9.299e-05	***
meatmeal	1	3389	3389	1.1266	0.2924	
soybean	1	56337	56337	18.7256	5.316e-05	***
sunflower	1	171	171	0.0567	0.8125	
Residuals	65	195556	3009			

ANOVA and GLM approaches match

ANOVA approach

Response: weight

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
feed	5	231129	46226	15.365	5.936e-10
Residuals	65	195556	3009		

GLM approach

Model used

Response: weight

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
horsebean	1	118991	118991	39.5510	3.064e-08
linseed	1	52241	52241	17.3640	9.299e-05
meatmeal	1	3389	3389	1.1266	0.2924
soybean	1	56337	56337	18.7256	5.316e-05
sunflower	1	171	171	0.0567	0.8125
Residuals	65	195556	3009		

Results presented

Response: weight

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sum(predictors)	5	231129	46226	15.365	5.936e-10
Residuals	65	195556	3009		

Testing differences between groups

- Some differences between mean are tested by coefficient tests...which?
- This is not all we want to know, though

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    323.583     15.834  20.436  < 2e-16 ***
feedhorsebean -163.383     23.485  -6.957 2.07e-09 ***
feedlinseed   -104.833     22.393  -4.682 1.49e-05 ***
feedmeatmeal   -46.674     22.896  -2.039 0.045567 *  
feedsoybean    -77.155     21.578  -3.576 0.000665 ***
feedsunflower    5.333     22.393   0.238 0.812495    
```

- We will learn about post-hocs in a GLM a little later

Different approaches to coding factors for GLM

- Dummy coding is a common approach, but there are others
- Choice of coding doesn’t affect the ANOVA table
- Interpretation of the coefficients changes
- Minimally, you should know that there are different choices, and be aware what your stat pack uses so you can interpret the coefficients correctly

Example: Deviation (effect) coding

Fertilizer means

$$\bar{x}_1 = 5.445$$
$$\bar{x}_2 = 3.999$$
$$\bar{x}_3 = 4.487$$

Grand mean is the baseline (intercept), not one of the levels

Slopes for all but the last group are differences from grand mean. The last group's difference from the grand mean is found by subtracting sum of other slopes.

Coefficients interpreted as differences from grand mean.

Useful when choice of baseline group is arbitrary. Also has some advantages for interpreting main effects and interactions.

This is how MINITAB does it, R can but not the default.

$$Y\hat{IE}LD = \begin{bmatrix} Fertil & Coeff \\ A & \alpha_1 \\ B & \alpha_2 \\ C & -\alpha_1 - \alpha_2 \end{bmatrix} + e$$

$$Y\hat{IE}LD = 4.6437 + \begin{bmatrix} Fertil & Coeff \\ A & 0.8103 \\ B & -0.6447 \\ C & -0.1566 \end{bmatrix}$$

Other coding systems:

- Difference coding – compares adjacent levels in an ordinal factor
  - Forward: level 1 vs 2, 2 vs 3, 3 vs 4
- Helmert coding – compares each level to the mean of subsequent levels
  - Compare 1 vs mean of 2,3,4; 2 vs mean of 3,4; 3 vs 4
- Orthogonal polynomial coding – tests for linear, quadratic, and cubic trends across ordinal levels (more later...)

What's the model?

Response?

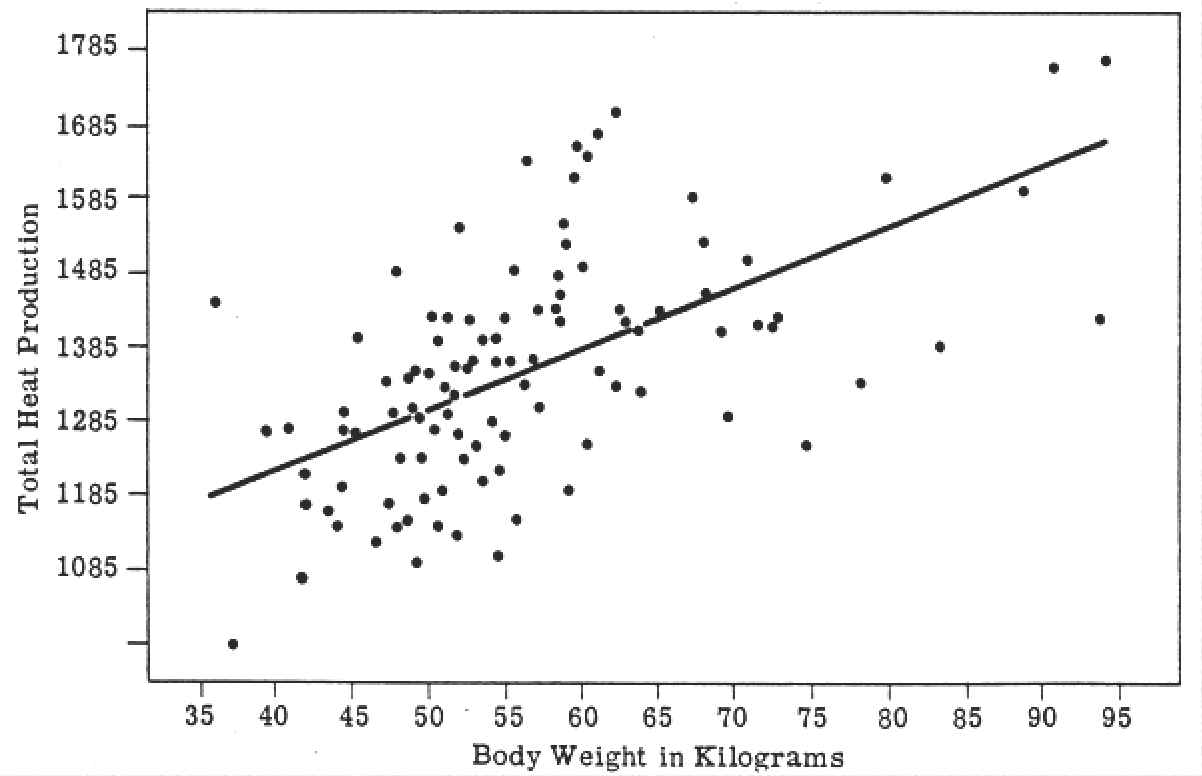
Predictor?

Would you expect an r^2 higher than 0.9?

What's the sign on the slope?

Is the intercept 0?

What would the regression equation look like?



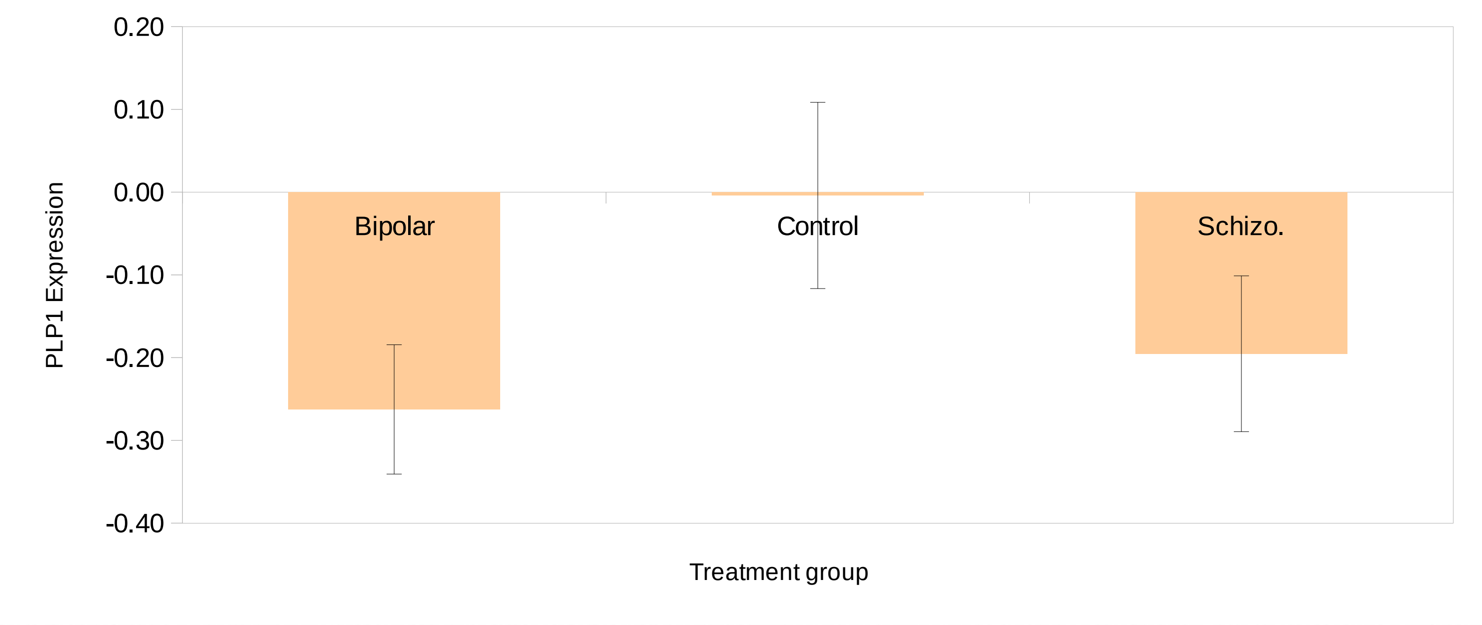
What's the model?

Response?

Predictor? Levels?

If Bipolar is the baseline group, what would the intercept represent?

What would the coefficient for Control represent?



Error bars are 2 se – do you expect these treatment groups to be significantly different?