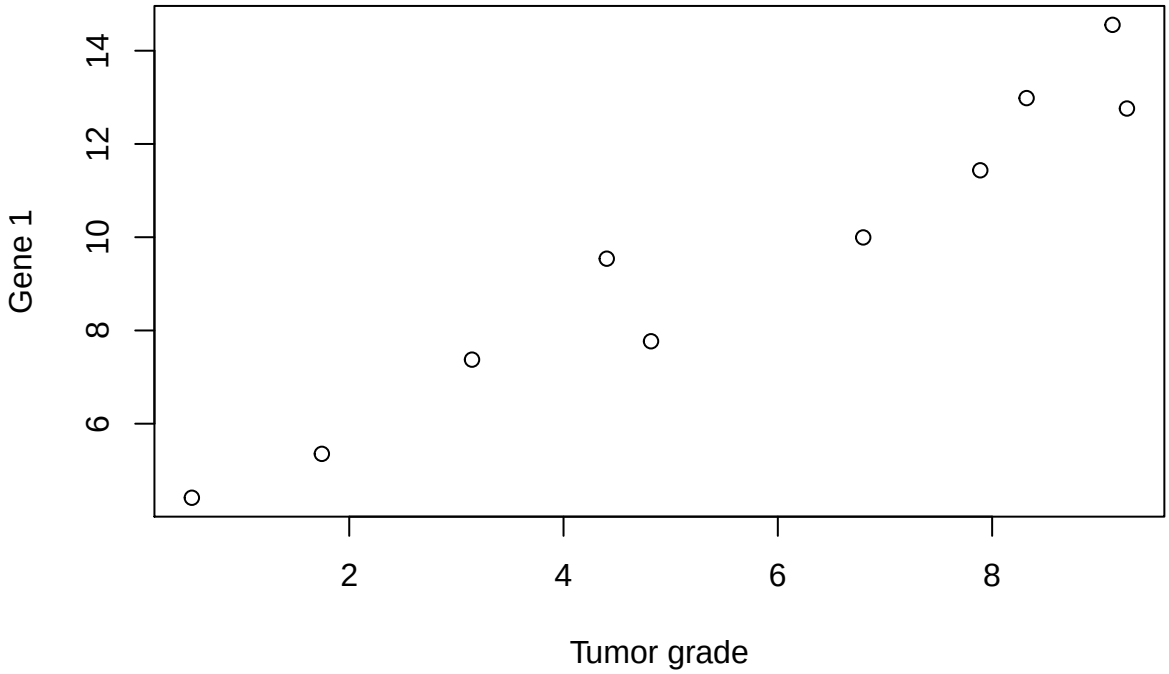
Design matrices

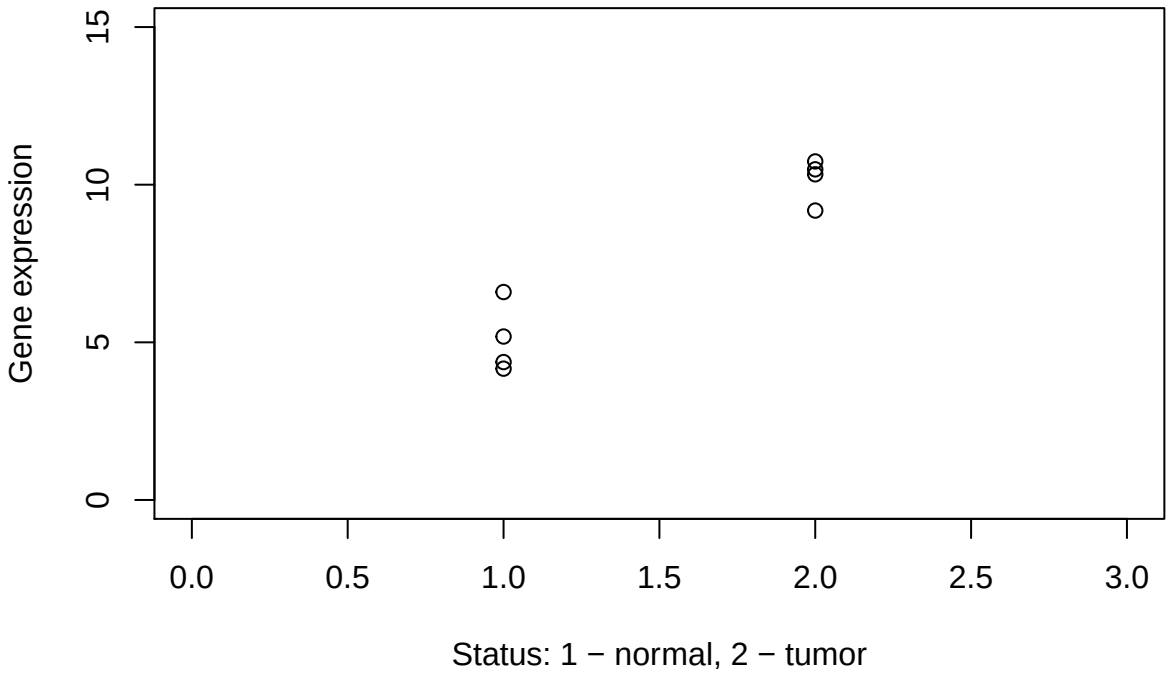
Linear regression overview



We measure tumor grade and expression of gene 1. We are interested in:

1) How useful was expression of gene 1 for predicting tumor grade? R^2
2) Was that relationship due to chance? $p - value$

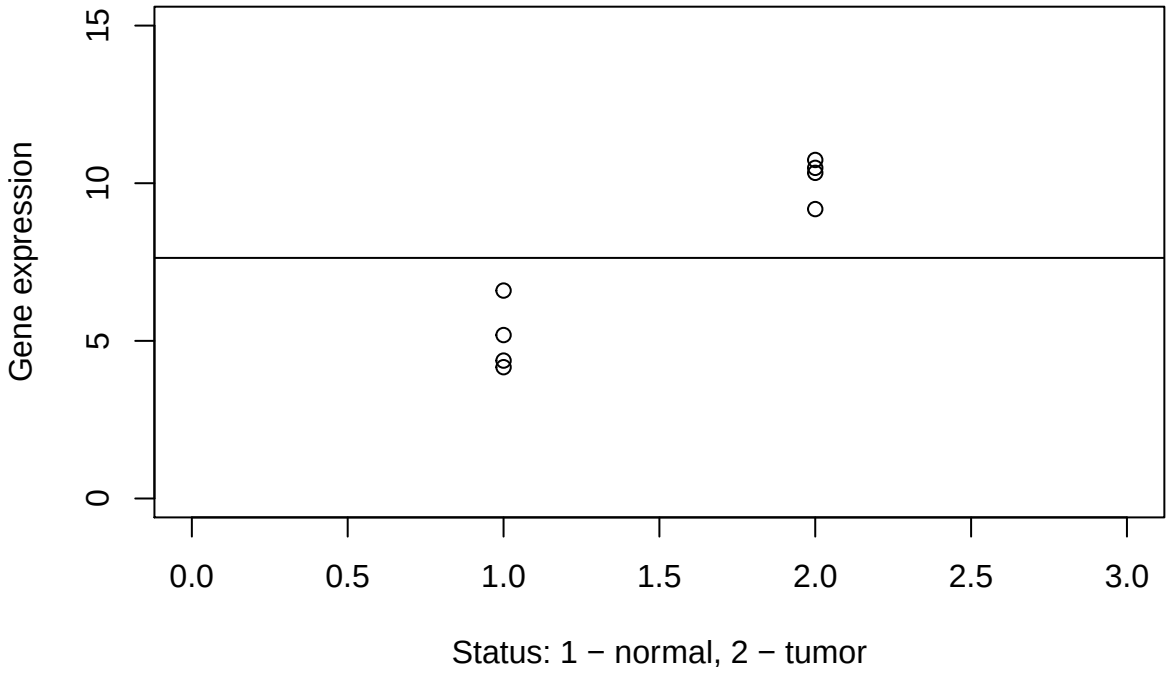
T-test



The goal of t-test is to compare meansand see if they are significantly different from each other.

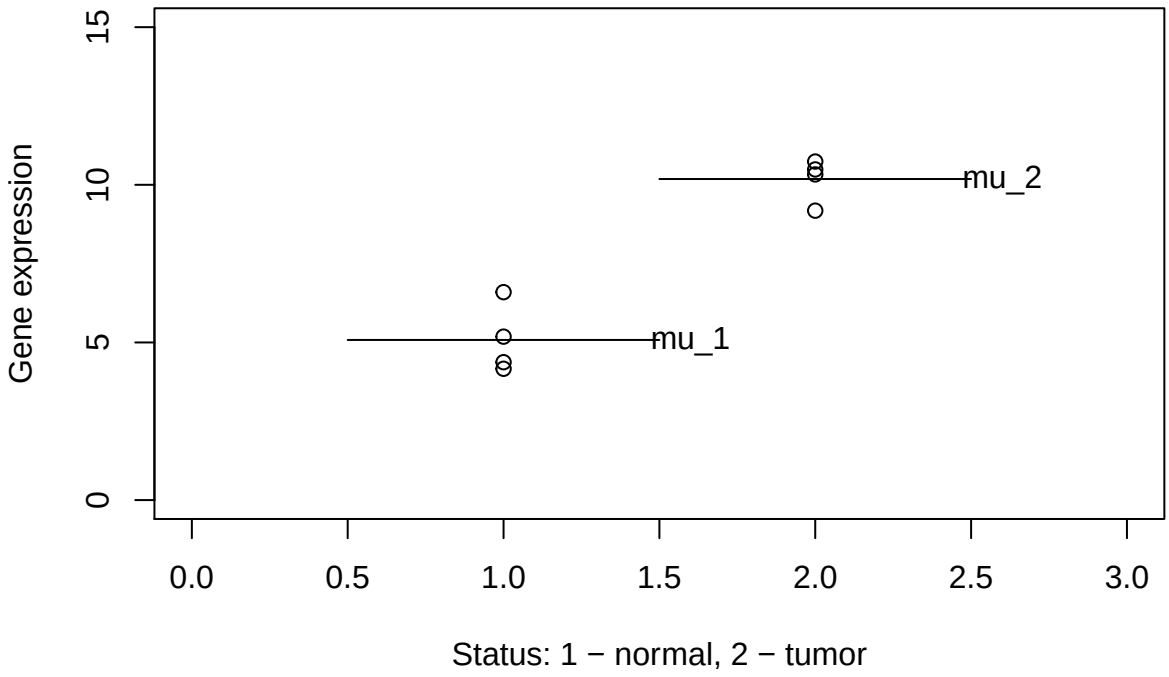
T-test in terms of linear regression

- Calculate the overall mean
- Calculate the sum of squared residuals around the mean SS_{mean}



- Fit a line to the data, separately for each group (mean = the least squares fit to the group of data)

- For each group, we can calculate SS_{fit}



- Combine two lines into a single equation. This will make the steps for computing F-statistics exactly the same as for the regression

```
# Combines both lines for the first group
y_11 = 1 * mu_1 + 0 * mu_2 + residual_11
y_12 = 1 * mu_1 + 0 * mu_2 + residual_12
y_13 = 1 * mu_1 + 0 * mu_2 + residual_13
y_14 = 1 * mu_1 + 0 * mu_2 + residual_14
# Combines both lines for the second group
y_21 = 0 * mu_1 + 1 * mu_2 + residual_21
y_22 = 0 * mu_1 + 1 * mu_2 + residual_22
y_23 = 0 * mu_1 + 1 * mu_2 + residual_23
y_24 = 0 * mu_1 + 1 * mu_2 + residual_24
```

- 1's and 0's serve as "switches" for each group. This is our design matrix X, one column per group.

```
1  0
1  0
1  0
1  0
0  1
0  1
0  1
0  1
```

- Get mu's and y's in a vector form, and

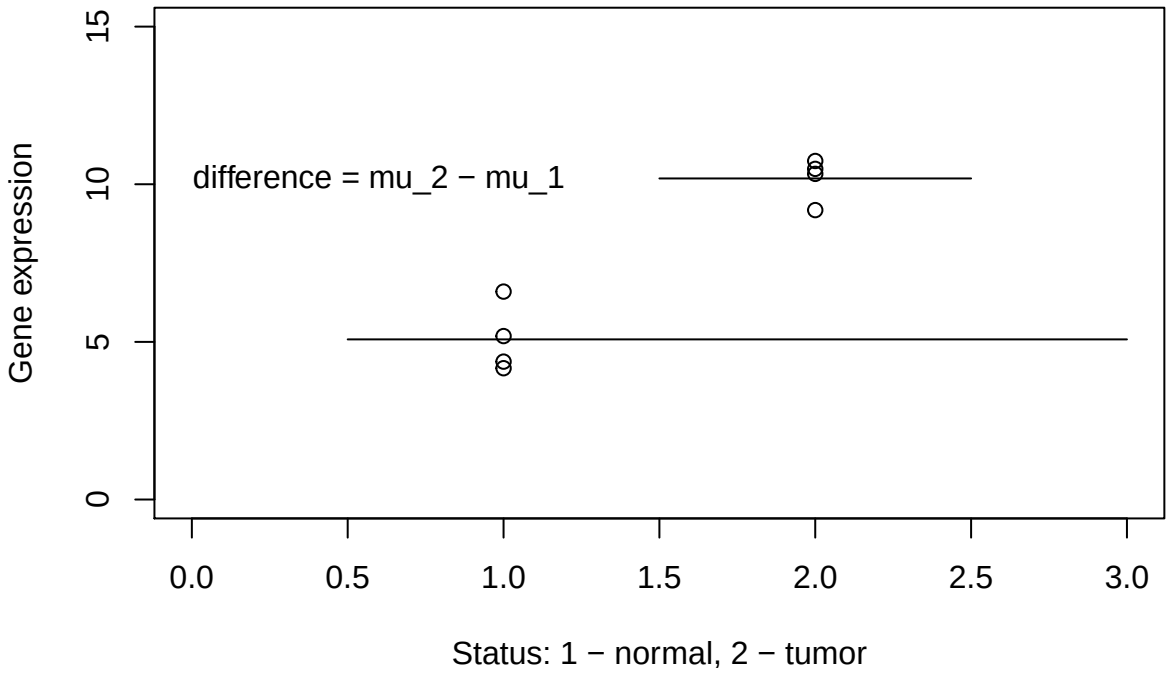
$$Y = X * MU + \epsilon$$

- Now, can calculate $F = \frac{(SS_{mean} - SS_{fit})/(p_{fit} - p_{mean})}{SS_{fit}/(n - p_{fit})}$ and the p-value
- p_{mean} - the number of parameters in the equation for the overall mean for gene expression (= 1)
- p_{fit} - the number of parameters in the line we fit for the data in individual groups (= 2)
- Same technique extends for multiple groups - ANOVA

A more common design matrix

```
1  0
1  0
1  0
1  0
1  1
1  1
1  1
1  1
```

- In this setup, all measurements contribute to the mean for the first group
- But only the measurements from the second group contribute to the *difference* between the first and the second group
- So the second column serves as a switch for the offset from the mean for the second group



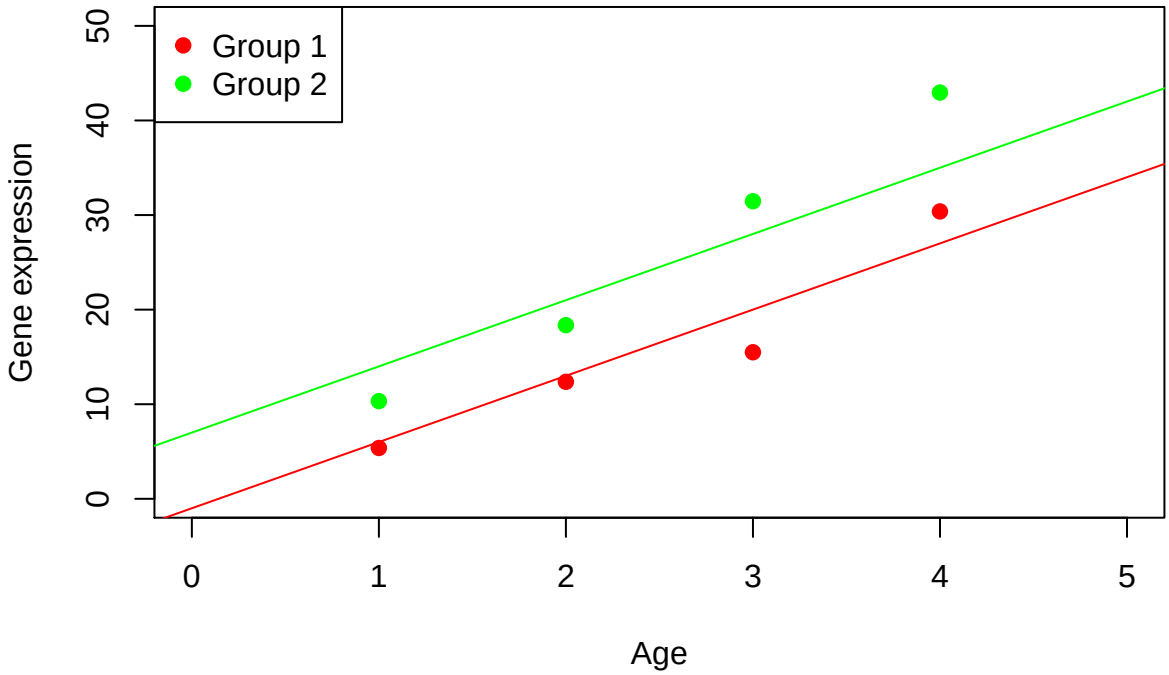

```
# Combines both lines for the first group
y_11 = 1 * mu_1 + 0 * difference_{mu_2 - mu_1} + residual_11
y_12 = 1 * mu_1 + 0 * difference_{mu_2 - mu_1} + residual_12
y_13 = 1 * mu_1 + 0 * difference_{mu_2 - mu_1} + residual_13
y_14 = 1 * mu_1 + 0 * difference_{mu_2 - mu_1} + residual_14
# Combines both lines for the second group
y_21 = 1 * mu_1 + 1 * difference_{mu_2 - mu_1} + residual_21
y_22 = 1 * mu_1 + 1 * difference_{mu_2 - mu_1} + residual_22
```

```
y_23 = 1 * mu_1 + 1 * difference_{mu_2 - mu_1} + residual_23
y_24 = 1 * mu_1 + 1 * difference_{mu_2 - mu_1} + residual_24
```

- Same way to calculate SS_{mean} and SS_{fit}
- Same number of equations
- Same number of parameters

Power of design matrices

- Say, in addition to group 1 and group 2, you have age variable.



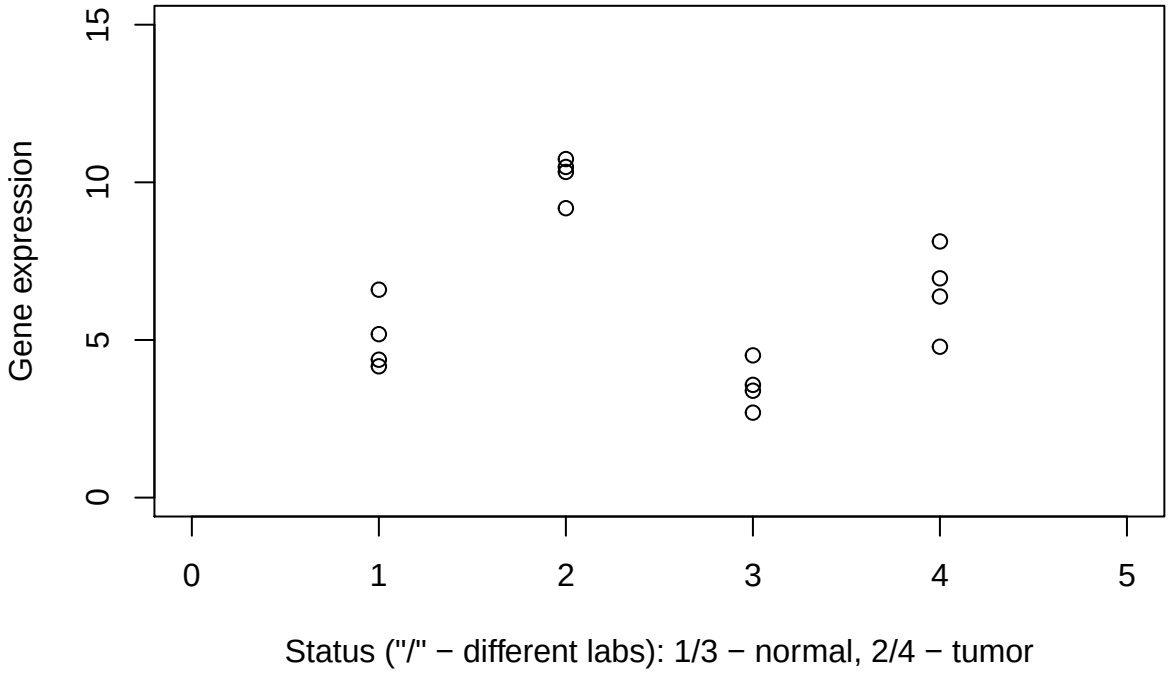
- We need to expand our model like $y = group1\_intercept + group2\_offset + slope$ - full model
- So, in our design matrix, first columns of 1's mean that both lines intercept the Y-axis, and specify the intercept for group 1
- The second column indicates the offset of group 2 measures
- The third column is the Age variable for each group

```
1  0  1
1  0  2
1  0  3
1  0  4
1  1  1
1  1  2
1  1  3
1  1  4
```

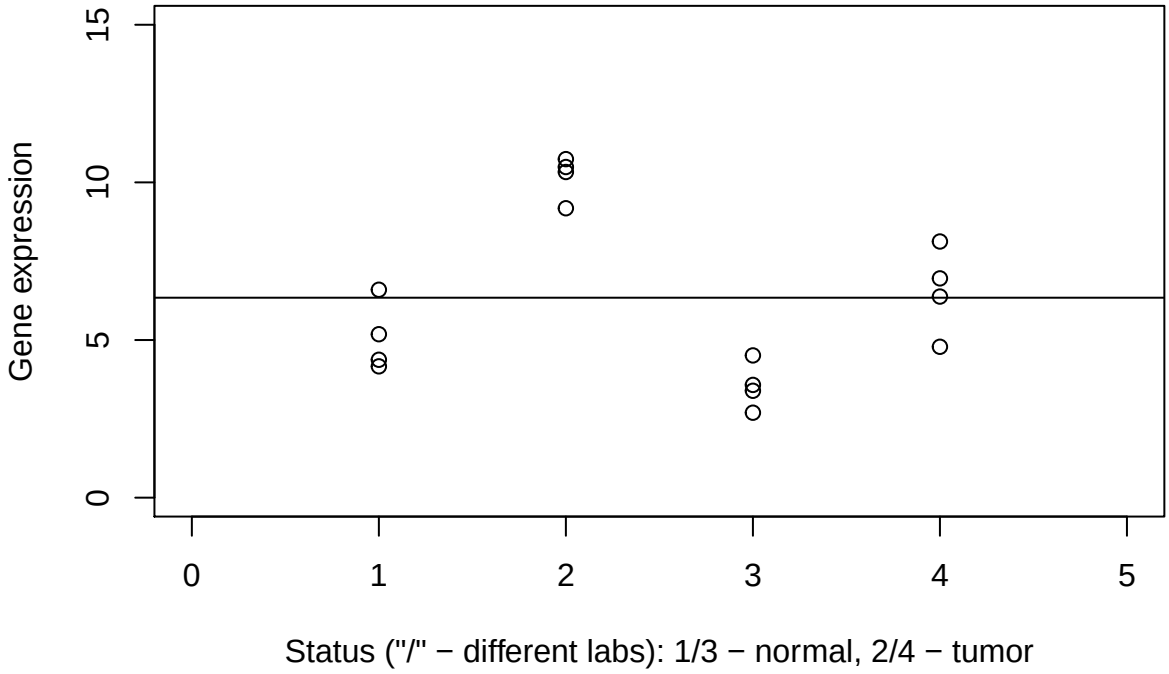
- Compare with the simple model $y = overall\_mean$
- Calculate how much better is the full model: $F = \frac{(SS_{simple} - SS_{full})/(p_{full} - p_{simple})}{SS_{full}/(n - p_{simple})}$

Batch effect

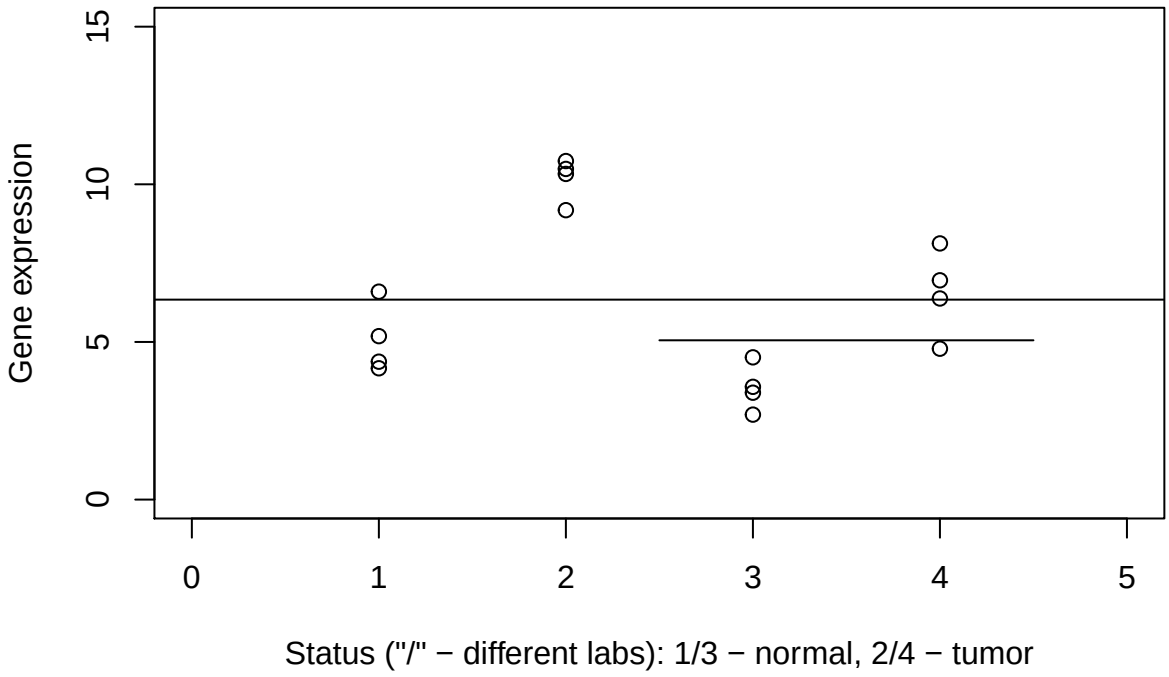
- Suppose you have measurements from two labs



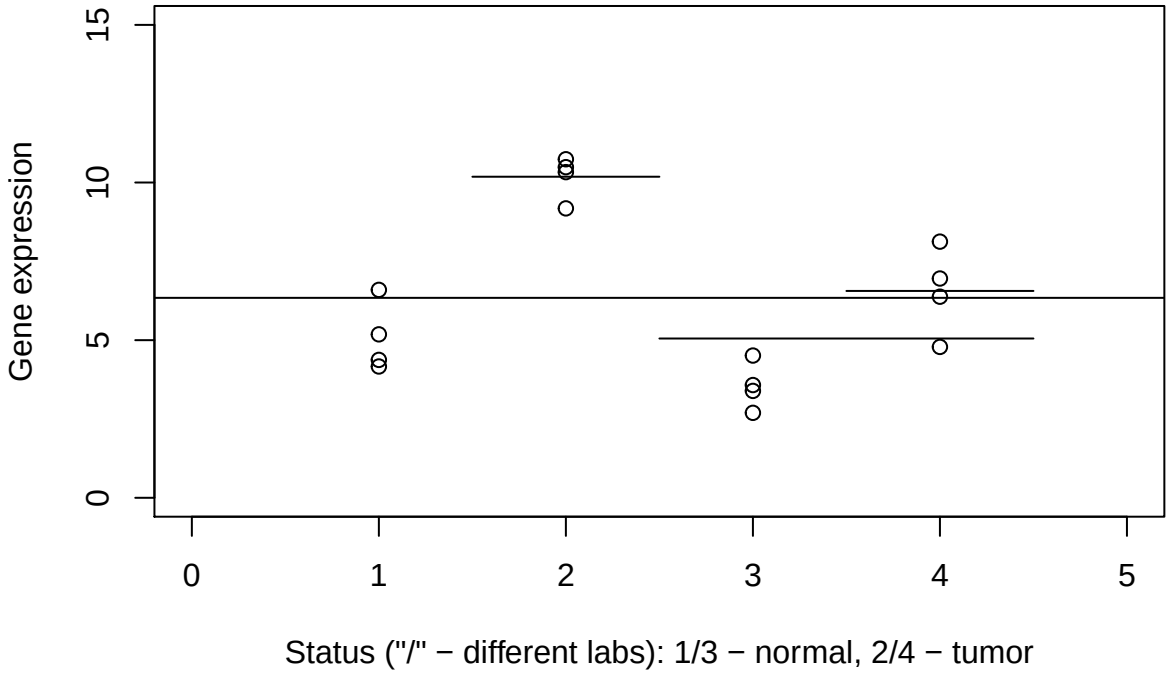
- First, add a term for the first lab normal group mean



- Second, add a term for the offset in measurements by the second lab



- Third, add a term for the offset of the tumor measurements



- The final model $y = lab1\_normal\_mean + lab2\_offset + difference_{tumor-normal}$, and the design matrix

```
1  0  0
1  0  0
1  0  0
1  0  0
```

```
1  0  1
1  0  1
1  0  1
1  0  1
1  1  0
1  1  0
1  1  0
1  1  0
1  1  1
1  1  1
1  1  1
1  1  1
```

- Does the lab effect matter? Compare the final model with a simpler one $y = lab1\_normal\_mean + difference_{tumor-normal}$

Learn more

- Law, Charity W., Kathleen Zeglinski, Xueyi Dong, Monther Alhamdoosh, Gordon K. Smyth, and Matthew E. Ritchie. "A Guide to Creating Design Matrices for Gene Expression Experiments." F1000Research (December 10, 2020) - Design matrices for various experimental designs. Means model or mean-reference model.
- Soneson, C, F Marini, F Geier, MI Love, and MB Stadler. "ExploreModelMatrix: Interactive Exploration for Improved Understanding of Design Matrices and Linear Models in R" F1000Research, (June 4, 2020).