

# **Statistical Methods for High Dimensional Biology**

## **STAT/BIOF/GSAT 540**

Lecture 7 – Linear models

Sara Mostafavi  
January 24 2018

**\*\*based on slides from Dr. Jenny Bryan\***

# Announcements

- Project groups – form a group today in the seminar.

Linear Models with R by Julian Faraway, Chapman & Hall/CRC Texts in Statistical Science, 2004.

One can find a related “eBook” or “PDF book” -- entitled “Practical Regression and Anova using R” -- in various places on the web. It seems to be an earlier, but very mature draft of the official book. Ingo Ruczinski provides a nice page for accessing each chapter as a PDF [here](#).

\*\* [www.biostat.jhsph.edu/~iruczins/teaching/jf/faraway.html](http://www.biostat.jhsph.edu/~iruczins/teaching/jf/faraway.html) \*\*

Applied Linear Statistical Models by Neter, Kutner, Nachtsheim, Wasserman. 4th ed, Irwin, 1996. (There is a more recent 5th edition.)

Venables WN, Ripley BD (2002) Modern applied statistics with S. Springer.

An Introduction to R (an “official” R document)

- Recall for ANOVA with devStage as categorical, you are treating each devStage \*independently\* without making any assumption about continuity across time..
- For factors with many levels, this results in estimation of too many parameters which detract from model interpretability and goals (e.g., you want to know if devStage matters in general)
  - You can get a bit of interpretability back by using F-test to assess effect across many levels jointly (R → ANOVA tools)

devStage	E16	P2	P6	P10	4_weeks
gType					
wt	$\theta$	$\beta_{P2}$	$\beta_{P6}$	$\beta_{P10}$	$\beta_{4\_weeks}$
NrIKO	$\tau_{NrIKO}$	$(\tau\beta)_{NrIKO,P2}$	$(\tau\beta)_{NrIKO,P6}$	$(\tau\beta)_{NrIKO,P10}$	$(\tau\beta)_{NrIKO,4\_weeks}$

# linear model style inferential output ... too granular?

Call:

```
lm(formula = prMat ~ gType * devStage)
```

Response[21567]: 1448159\_at

Residuals:

Min	Q1	Median	Q3	Max
-0.2725	-0.0735	0.0025	0.0955	0.2163

Coefficients:

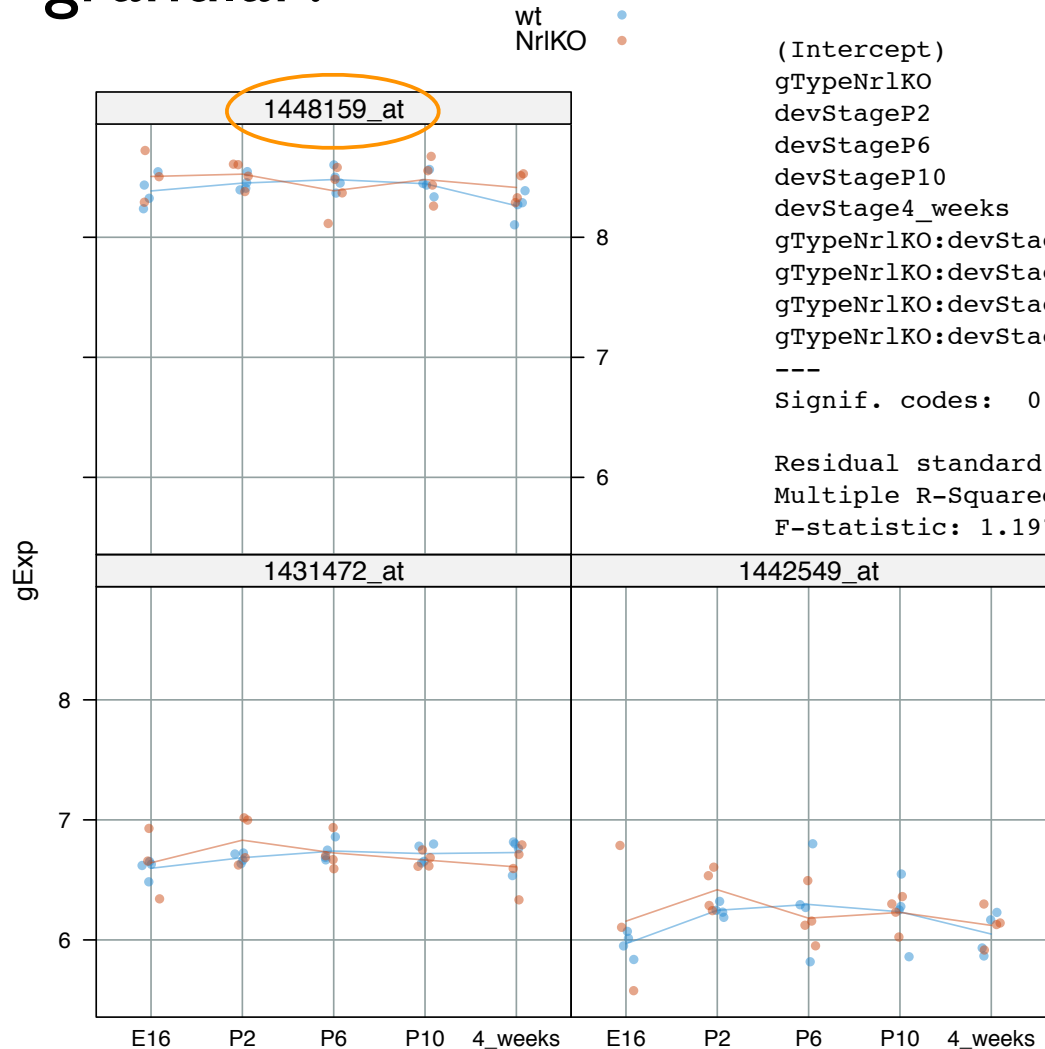
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	8.38600	0.06903	121.475	<2e-16 ***
gTypeNrlKO	0.12067	0.10545	1.144	0.262
devStageP2	0.06550	0.09763	0.671	0.508
devStageP6	0.09500	0.09763	0.973	0.339
devStageP10	0.06050	0.09763	0.620	0.540
devStage4_weeks	-0.12300	0.09763	-1.260	0.218
gTypeNrlKO:devStageP2	-0.04617	0.14371	-0.321	0.750
gTypeNrlKO:devStageP6	-0.21417	0.14371	-1.490	0.147
gTypeNrlKO:devStageP10	-0.08617	0.14371	-0.600	0.553
gTypeNrlKO:devStage4_weeks	0.03133	0.14371	0.218	0.829

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

Residual standard error: 0.1381 on 29 degrees of freedom

Multiple R-Squared: 0.2709, Adjusted R-squared: 0.04463

F-statistic: 1.197 on 9 and 29 DF, p-value: 0.3339



# two-way ANOVA

## style inferential

## output ... too

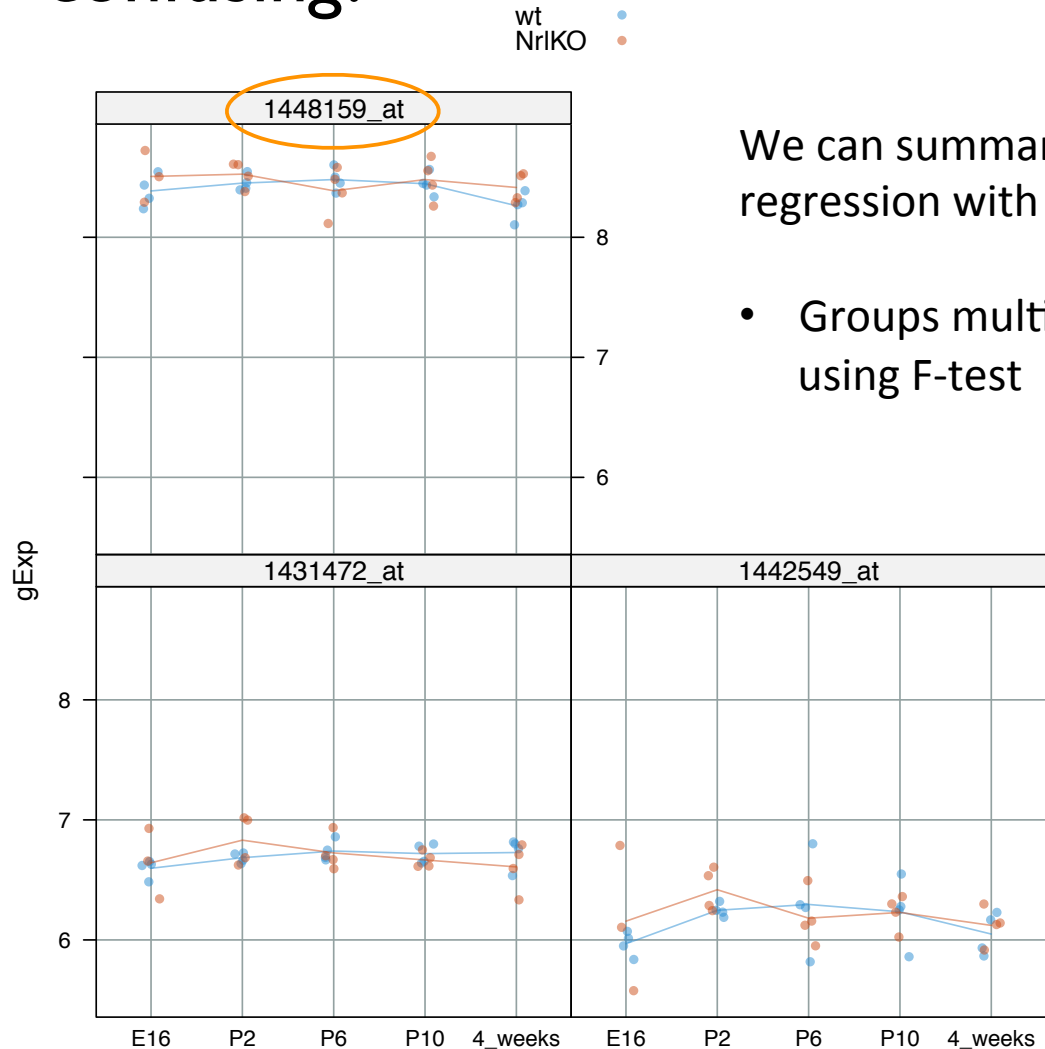
## confusing?

```
> anova(lm(gExp ~ gType * devStage, jDat))
```

Analysis of Variance Table

Response: gExp

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gType	1	0.02985	0.029848	1.5657	0.2208
devStage	4	0.10365	0.025914	1.3594	0.2722
gType:devStage	4	0.07191	0.017977	0.9430	0.4532
Residuals	29	0.55283	0.019063		



We can summarize the result of multi-level linear regression with categorical variable using ANOVA tools.

- Groups multiple levels and assess significance jointly using F-test

Quick note: order of main effect matters in unbalanced design

You can use “ANOVA” in R to address this

```
> anova(lm(gExp ~ gType * devStage, jDat))
Analysis of Variance Table
```

Response: gExp

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gType	1	0.02985	0.029848	1.5657	0.2208
devStage	4	0.10365	0.025914	1.3594	0.2722
gType:devStage	4	0.07191	0.017977	0.9430	0.4532
Residuals	29	0.55283	0.019063		

```
> anova(lm(gExp ~ devStage * gType, jDat))
Analysis of Variance Table
```

Response: gExp

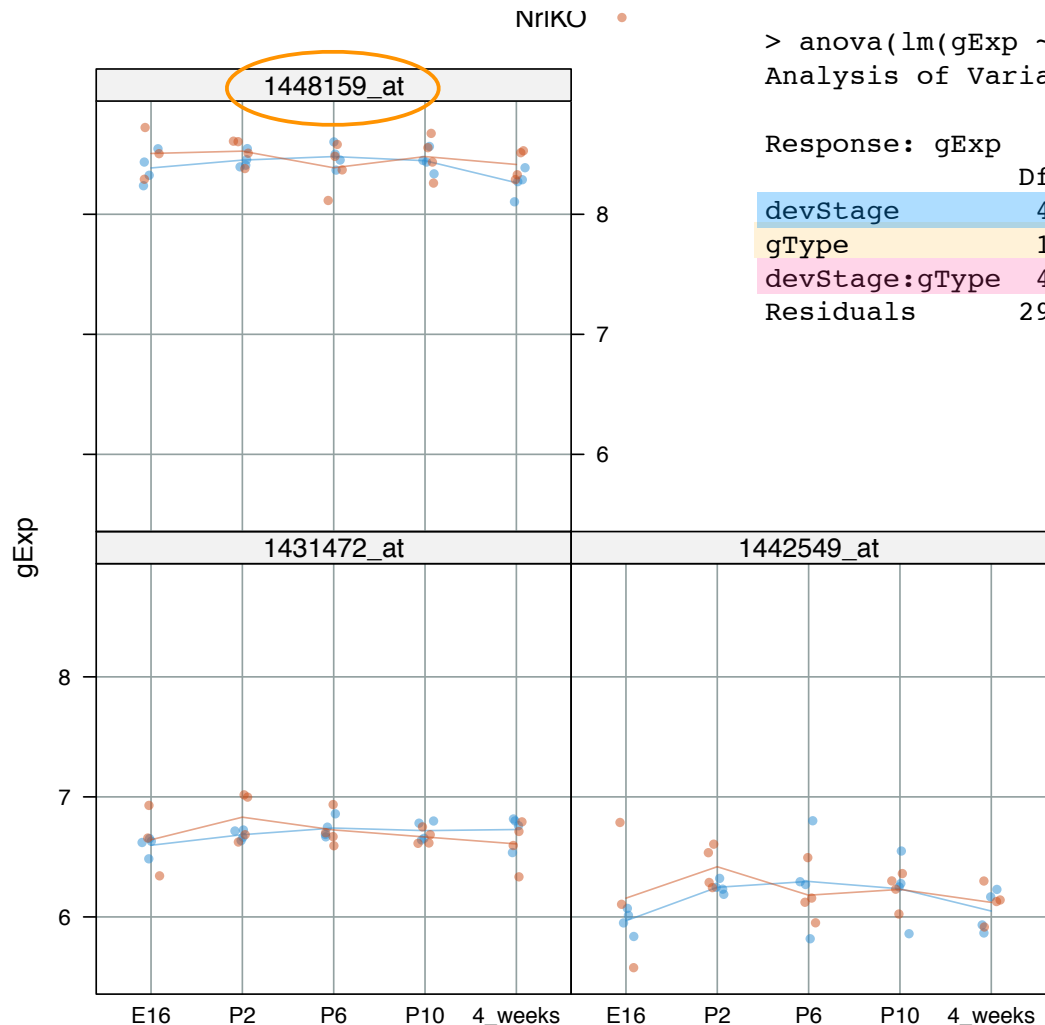
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
devStage	4	0.10328	0.025819	1.3544	0.2739
gType	1	0.03022	0.030225	1.5855	0.2180
devStage:gType	4	0.07191	0.017977	0.9430	0.4532
Residuals	29	0.55283	0.019063		

ANOVA tables address whether, e.g., all the interaction effects, are non-zero

note the agreement above for the interaction gType:devStage

note the discrepancies above for main effects ... depends on order ... related to the sequential nature of Type I sums of squares

we are suffering for our unbalanced design :(



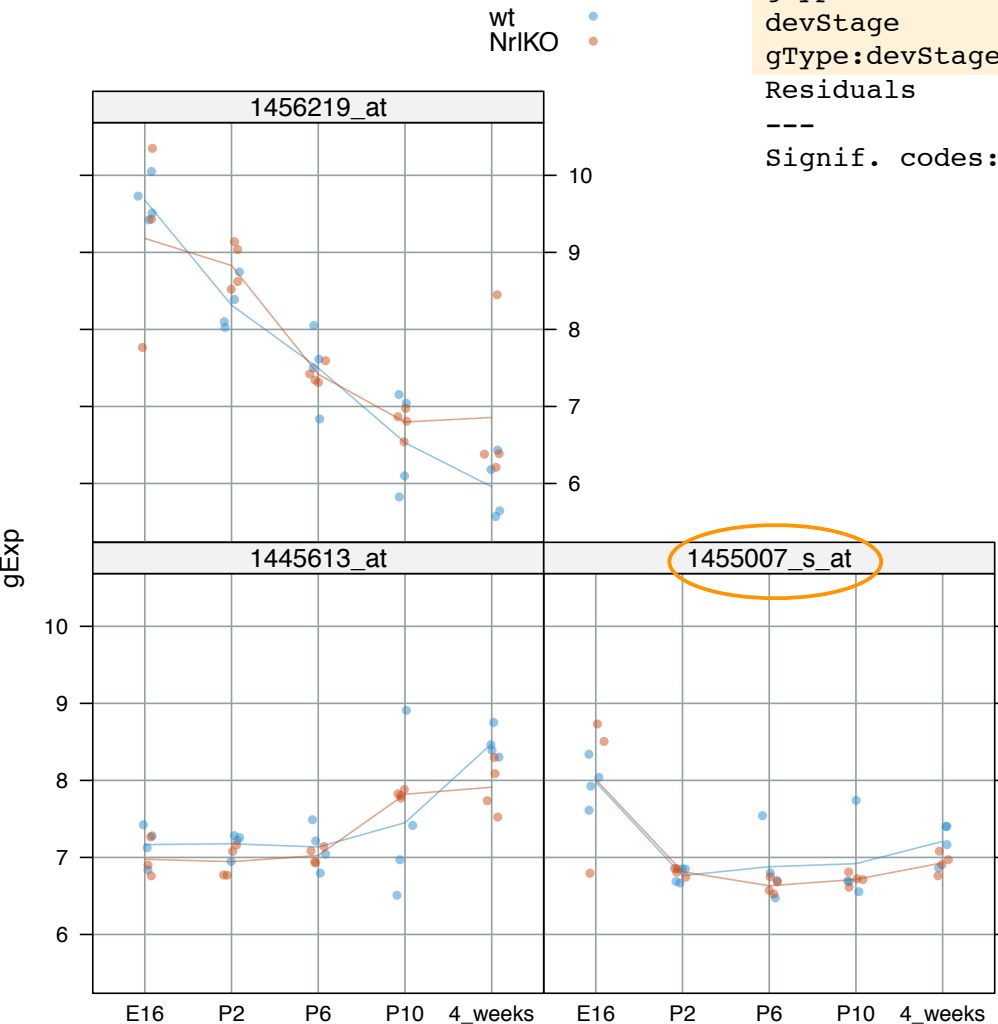


# Analysis of Variance Table

-----  
Response[26301]: 1455007\_s\_at

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gType	1	0.3209	0.32092	2.1120	0.1569
devStage	4	7.7431	1.93578	12.7394	4.204e-06 ***
gType:devStage	4	0.1927	0.04818	0.3171	0.8642
Residuals	29	4.4066	0.15195		

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



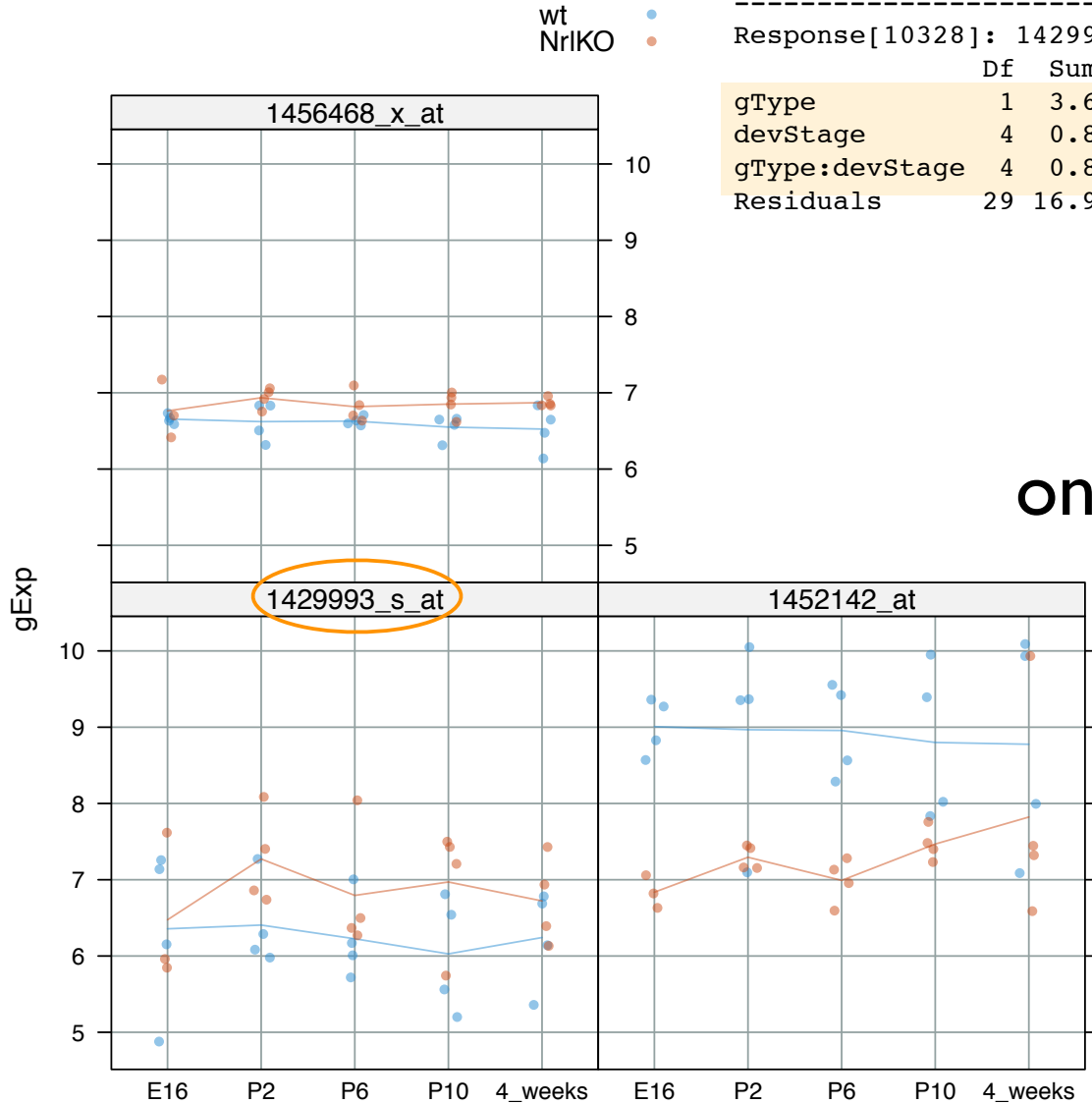
only devStage matters

# Analysis of Variance Table

Response[10328]: 1429993\_s\_at

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gType	1	3.6819	3.6819	6.3094	0.01783 *
devStage	4	0.8028	0.2007	0.3439	0.84603
gType:devStage	4	0.8034	0.2008	0.3442	0.84586
Residuals	29	16.9231	0.5836		

only gType matters



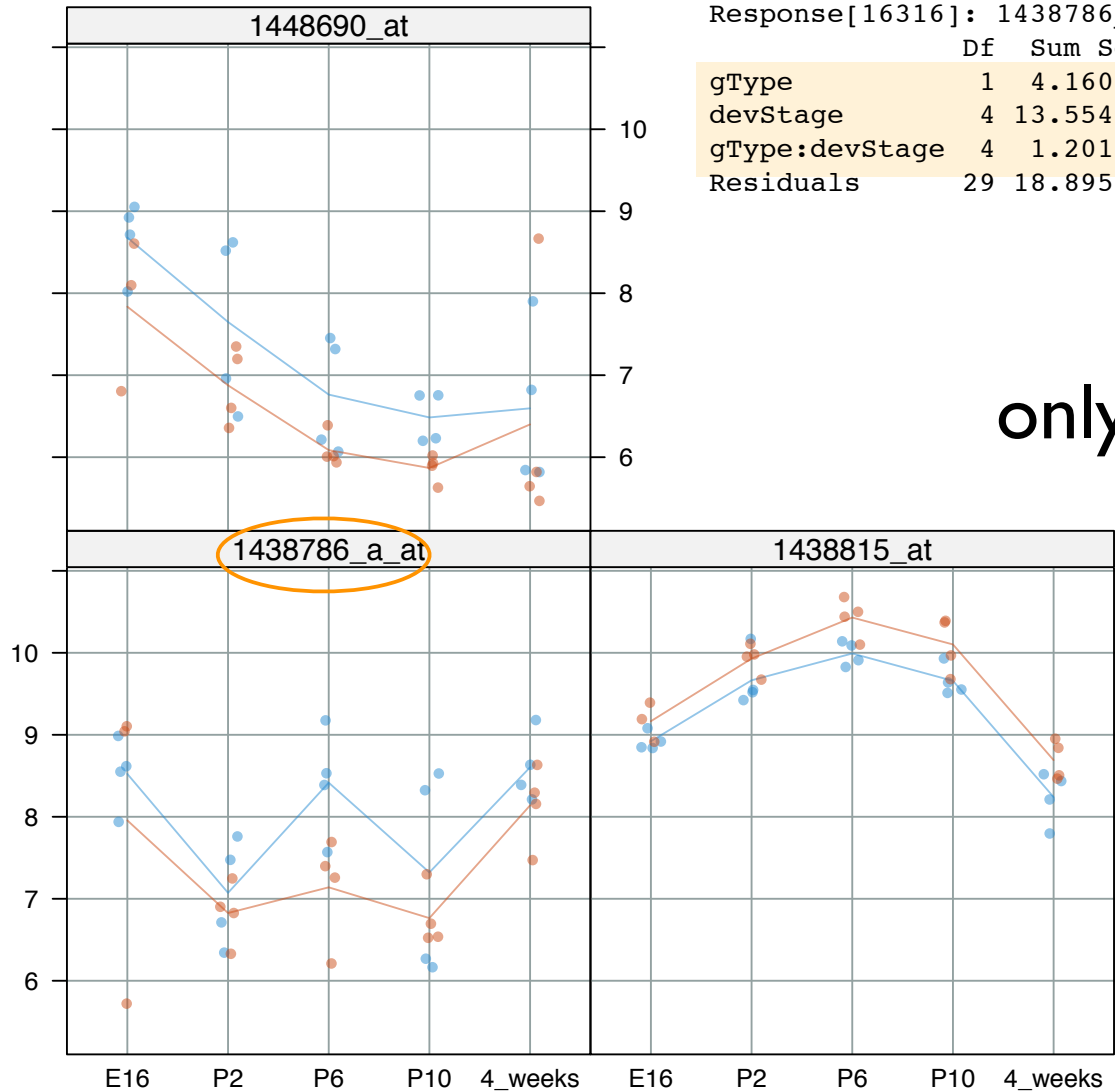
wt      ● Analysis of Variance Table  
 Nr1KO    ●

-----  
 Response[16316]: 1438786\_a\_at

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
gType	1	4.1606	4.1606	6.3855	0.017216	*
devStage	4	13.5545	3.3886	5.2008	0.002774	**
gType:devStage	4	1.2014	0.3003	0.4610	0.763712	
Residuals	29	18.8953	0.6516			

only main effects

gExp



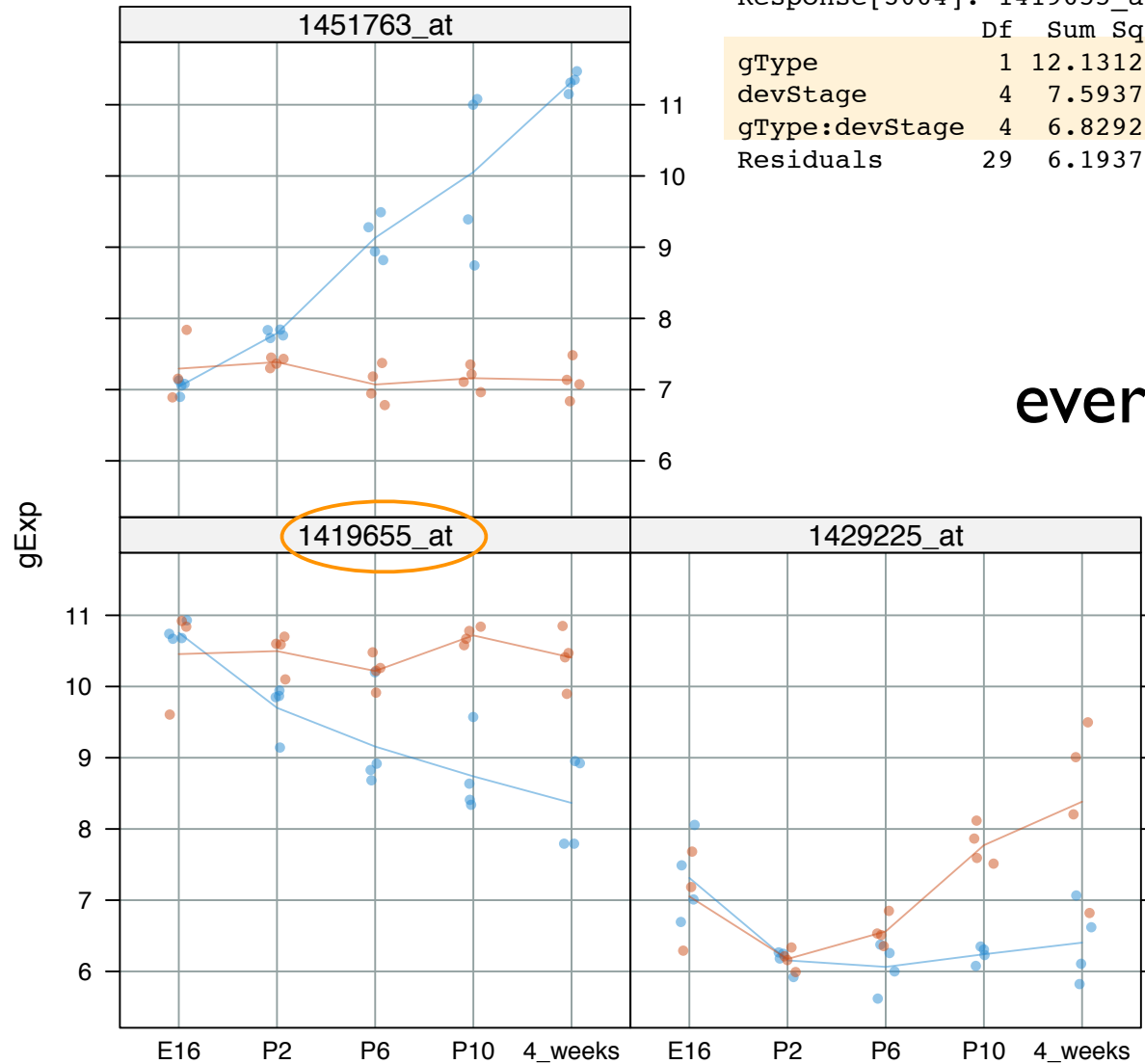
# Analysis of Variance Table

wt  
NrIKO

Response[3064]: 1419655\_at

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
gType	1	12.1312	12.1312	56.8008	2.623e-08	***
devStage	4	7.5937	1.8984	8.8888	8.210e-05	***
gType:devStage	4	6.8292	1.7073	7.9939	0.0001798	***
Residuals	29	6.1937	0.2136			

everything's going on



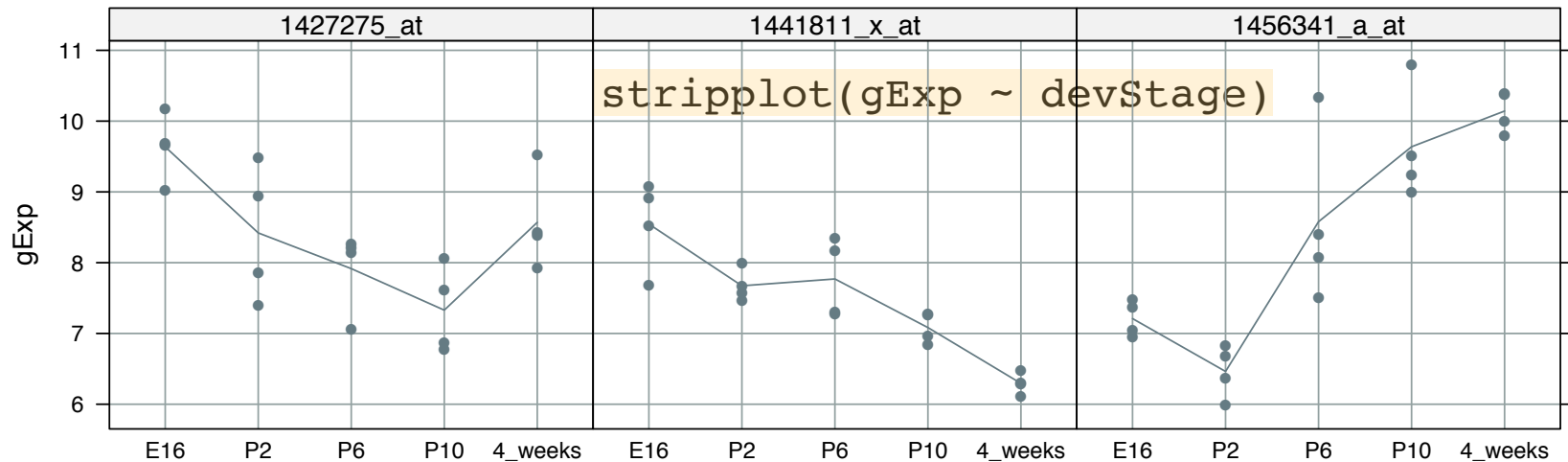
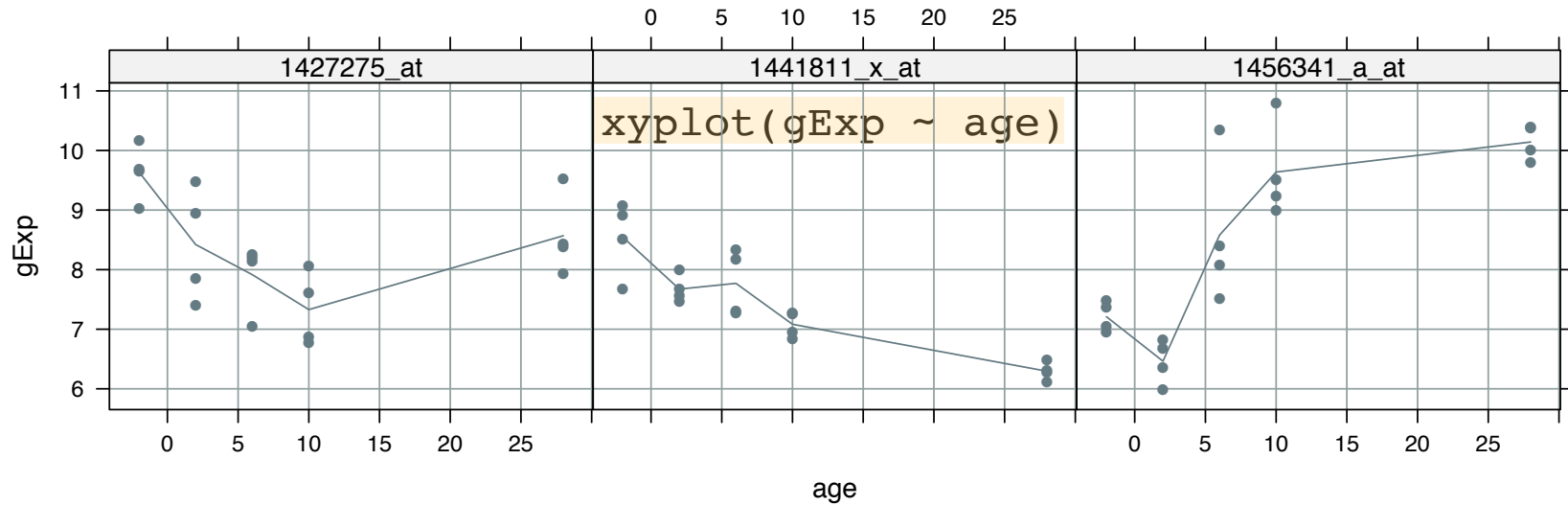
- Recall for ANOVA with devStage as categorical, you are treating each devStage \*independently\* without making any assumption about continuity across time..
- Thus, you are not using any knowledge about continuity to rule out nonsensical patterns
- For factors with many levels, this results in estimation of too many parameters which detract from model interpretability and goals (e.g., you want to know if devStage matters in general)

devStage gType	E16	P2	P6	P10	4_weeks
wt	$\theta$	$\beta_{P2}$	$\beta_{P6}$	$\beta_{P10}$	$\beta_{4\_weeks}$
NrIKO	$\tau_{NrIKO}$	$(\tau\beta)_{NrIKO,P2}$	$(\tau\beta)_{NrIKO,P6}$	$(\tau\beta)_{NrIKO,P10}$	$(\tau\beta)_{NrIKO,4\_weeks}$

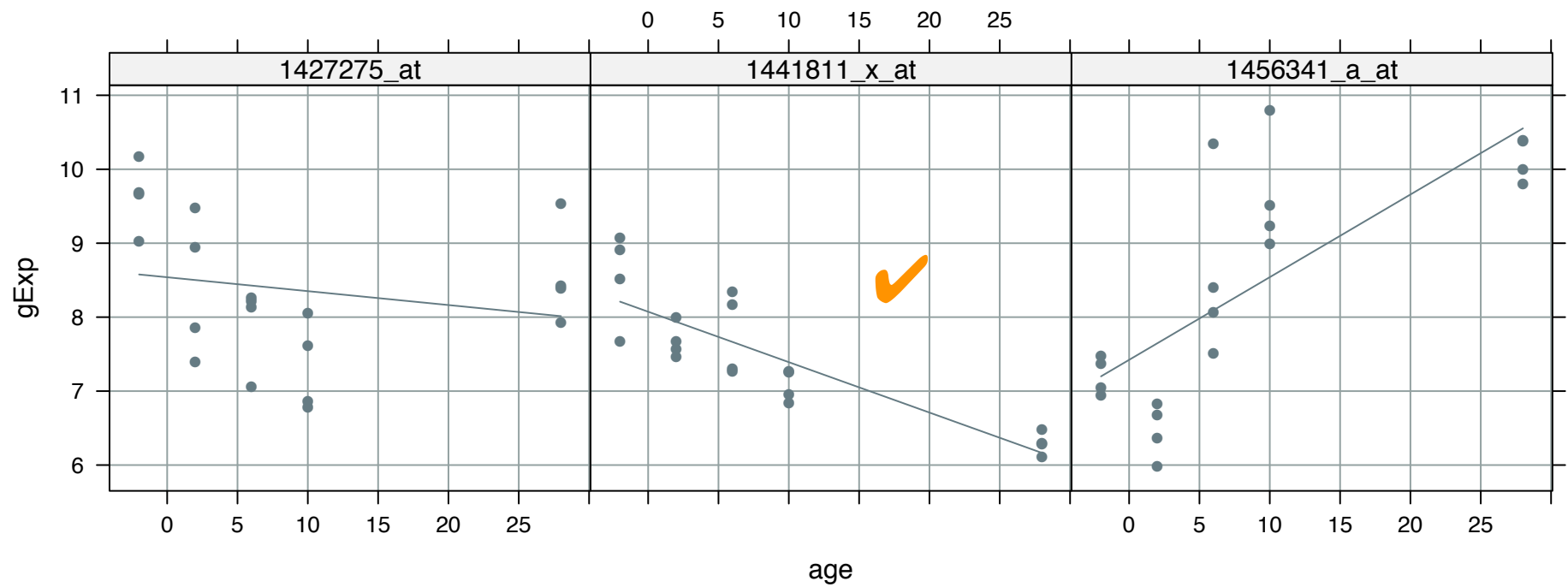
# Linear model – linear regression

- Provide a general framework for modeling the relationship between response variable and some explanatory variables (factors/features/covariates).
- Anova – provides a set of tools for analyzing the results of a linear regression model, when the independent variables are categorical.
- When dealing with factors with many levels, its typically a better approach to translate them continuous factors and hence use the generic engine of linear regression.

for starters, let's just work with wild type data for 3 example probesets



Kind of a different look to the data, no?



linear looks reasonable for 1, but  
not the other two



- For now, we'll just assume a linear fit is good enough. We'll come back to relax this later.
- Under Occam's Razor, you should always assume the simplest model (e.g., with least parameters/DoF) unless *statistically* proven otherwise.

```

> ## recode() is from add-on package 'car'

> prDes$age <-
+   recode(prDes$devStage,
+         "'E16'=-2; 'P2'=2; 'P6'=6; 'P10'=10; '4_weeks'=28",
+         as.factor.result = FALSE)

> peek(prDes)
      sample devStage gType age
Sample_22      22      E16   wt  -2
Sample_16      16      E16 NrlKO -2
Sample_5         5       P2 NrlKO  2
Sample_31      31       P6   wt   6
Sample_15      15      P10 NrlKO 10
Sample_36      36  4_weeks   wt  28
Sample_2         2  4_weeks NrlKO 28

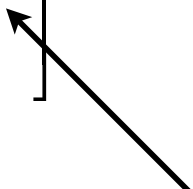
> str(prDes)
'data.frame':  39 obs. of  4 variables:
 $ sample  : num  20 21 22 23 16 17 6 24 25 26 ...
 $ devStage: Factor w/ 5 levels "E16","P2","P6",...: 1 1 1 1 1 1 1 2 2 2 ...
 $ gType   : Factor w/ 2 levels "wt","NrlKO": 1 1 1 1 2 2 2 1 1 1 ...
 $ age     : num  -2 -2 -2 -2 -2 -2 -2 -2 2 2 2 ...

```

meet our new quantitative covariate or predictor ...  
*age*, which is a new version of the factor *devStage*

# Plain vanilla linear model, matrix formulation

$$Y = X\alpha + \varepsilon$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \alpha = \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$


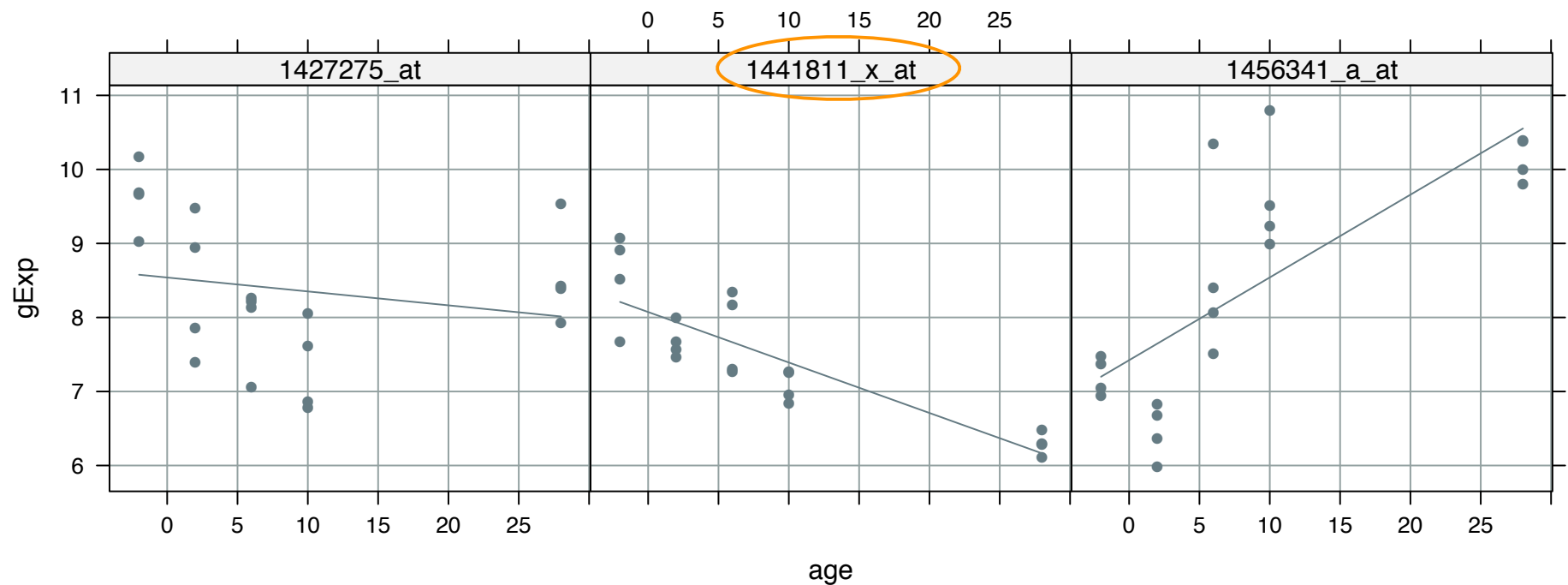
Here's what a design matrix would look like with 1 quantitative covariate.

$$Y = X\alpha + \varepsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \alpha_0 \cdot 1 + \alpha_1 \cdot x_1 \\ \alpha_0 \cdot 1 + \alpha_1 \cdot x_2 \\ \vdots \\ \alpha_0 \cdot 1 + \alpha_1 \cdot x_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \alpha_0 + \alpha_1 x_1 + \varepsilon_1 \\ \alpha_0 + \alpha_1 x_2 + \varepsilon_2 \\ \vdots \\ \alpha_0 + \alpha_1 x_n + \varepsilon_n \end{bmatrix}$$

$$y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i$$

Remember / convince yourself that the matrix algebra does indeed reproduce simple linear regression.



```
> summary(linFits[["1441811_x_at"]])
```

Call:

```
lm(formula = gExp ~ age, data = z)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.55059	-0.37459	-0.08398	0.31011	0.86827

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	8.073374	0.133118	60.648	< 2e-16 ***
age	-0.068179	0.009771	-6.978	1.62e-06 ***

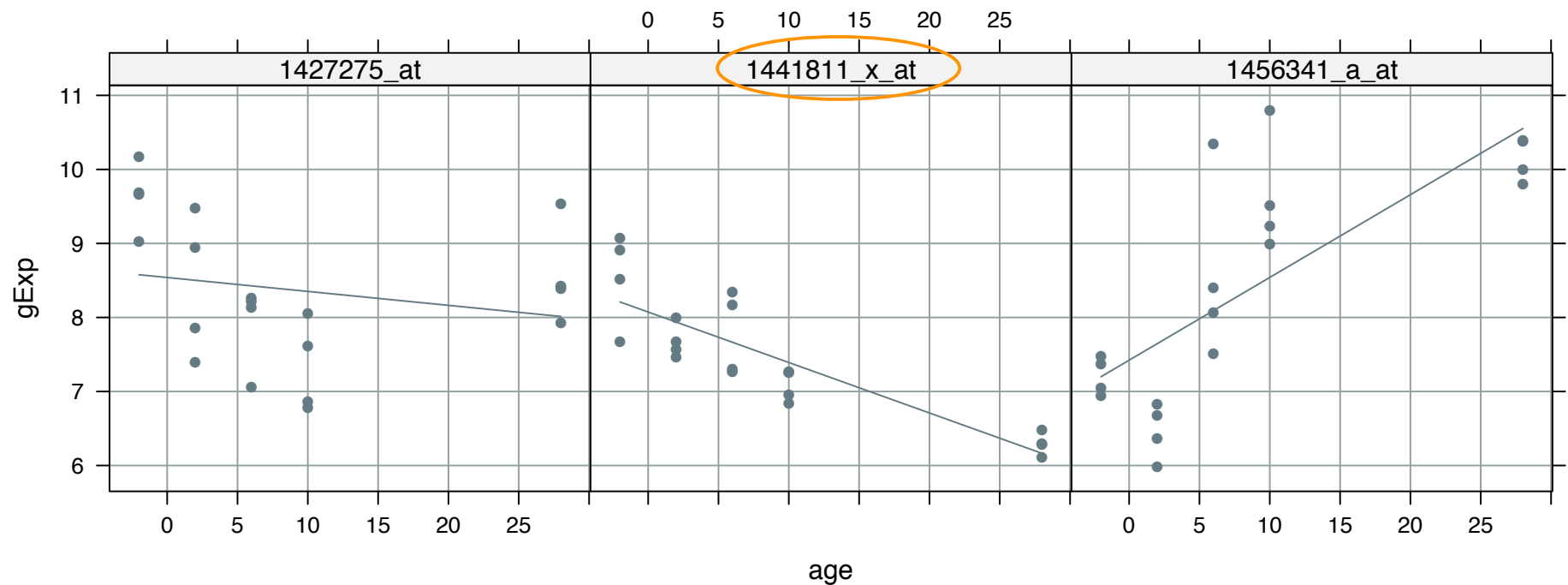
---

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

Residual standard error: 0.4545 on 18 degrees of freedom

Multiple R-squared: 0.7301, Adjusted R-squared: 0.7151

F-statistic: 48.69 on 1 and 18 DF, p-value: 1.622e-06



Linear regression framework is simple but very **\*\*powerful\*\***:

Now you have an equation for predicting the expression at any **\*arbitrary\*** age (not only those time points you measured)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	8.073374	0.133118	60.648	< 2e-16 ***
age	-0.068179	0.009771	-6.978	1.62e-06 ***

---

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

Residual standard error: 0.4545 on 18 degrees of freedom

Multiple R-squared: 0.7301, Adjusted R-squared: 0.7151

F-statistic: 48.69 on 1 and 18 DF, p-value: 1.622e-06

Now we have a simple framework for relating outcome to an arbitrary set of variables. But how do we \*estimate\* the parameters?

- How do we estimate the model parameters ( $\alpha$ 's)?
- It's the same setup for categorical and continuous factors
- Two ways of looking at it: equivalent results in terms of math
  1. Maximizing model likelihood (Probabilistic interpretation)
  2. Minimizing least squares error



# Maximum Likelihood Estimation (MLE)

1. Maximizing model likelihood (Probabilistic interpretation)
  - a. Write down likelihood of the data and simplify it based on IID assumption
  - b. Take log of the likelihood
  - c. Find the parameters that maximize the log of likelihood wrt to the unknown parameters

# Maximum Likelihood Estimation (MLE)

## 1. Maximizing model likelihood (Probabilistic interpretation)

$$\begin{aligned} p(\text{data}|\text{model}) &= p(y | X, \alpha_0, \alpha_{age}) \\ &= \prod_i N(y_i | \alpha_0 + \alpha_{age}, \sigma) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - (\alpha_0 + \alpha_{age}))^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y - X\alpha)^T (y - X\alpha)\right) \end{aligned}$$

# Maximum Likelihood Estimation (MLE)

## 1. Maximizing model likelihood (Probabilistic interpretation)

$$\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(y - X\alpha)^T(y - X\alpha)$$

$$\arg\max_{\alpha} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(y - X\alpha)^T(y - X\alpha)$$

$$= \arg\max_{\alpha} (y - X\alpha)^T(y - X\alpha)$$

$$\arg\max_{\alpha} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(y - X\alpha)^T(y - X\alpha)$$

$$\frac{\partial}{\partial \alpha} (y - X\alpha)^T(y - X\alpha) = 0$$

$$X^T(y - X\alpha)^T = 0$$

$$X^T X\alpha = X^T y$$

$$\alpha = (X^T X)^{-1} X^T y$$

# Greatest Hits of Regression Results (normal iid errors)

$$Y = X\alpha + \varepsilon \quad \text{regression model}$$

$$\hat{\alpha} = (X^T X)^{-1} X^T Y \quad \text{the MLE and OLS estimator of } \alpha$$

$$\hat{Y} = X\hat{\alpha} \quad \text{the fitted or predicted values}$$

$$\hat{Y} = X(X^T X)^{-1} X^T Y = HY \quad \text{where } H = X(X^T X)^{-1} X^T \text{ is called the "hat matrix"}$$

$$\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\alpha} \quad \text{the residuals (note NOT the same as the errors } \varepsilon)$$

$$\hat{\sigma}^2 = \frac{1}{n-p} \hat{\varepsilon}^T \hat{\varepsilon} \quad \text{the estimated error variance (} p \text{ is the dimension of } \alpha)$$

$$\hat{V}(\hat{\alpha}) = \hat{\sigma}^2 (X^T X)^{-1} \quad \text{the estimated covariance matrix of } \hat{\alpha}$$

estimated standard errors for the estimated regression coefficients --  $\widehat{se}(\hat{\alpha}_j)$  --  
are obtained by taking the square root of the diagonal elements of  $\hat{V}(\hat{\alpha})$

# Inference in Regression (normal iid errors)

$Y = X\alpha + \varepsilon$  regression model

$\hat{\alpha} = (X^T X)^{-1} X^T Y$  the MLE and OLS estimator of  $\alpha$

$\hat{\sigma}^2 = \frac{1}{n-p} \hat{\varepsilon}^T \hat{\varepsilon}$  the estimated error variance

$\hat{V}(\hat{\alpha}) = \hat{\sigma}^2 (X^T X)^{-1}$  the estimated covariance matrix of  $\hat{\alpha}$

How test  $H_0 : \alpha_j = 0$ ?

With a t-statistic. Under  $H_0$ , we have (at least approximately) that:

$$\frac{\hat{\alpha}_j}{\widehat{se}(\hat{\alpha}_j)} \sim t_{n-p}$$

so a p-value is obtained by computing a tail probability for the observed value of  $\hat{\alpha}_j$  from a  $t_{n-p}$  distribution.

Increasing the complexity of our linear regression example

1 quantitative and 1 categorical variable

Age and genotype (WT vs KO)

$$y_{ij} = \alpha_{0,wt} + \tau_{0,j} + (\alpha_{1,wt} + \tau_{1,j})age_i + \varepsilon_{ij}$$

where  $j \in \{wt, NrlKO\}$

$$i = 1, 2, \dots, n_j$$

$$\tau_{0,wt} = \tau_{1,wt} \equiv 0$$

```
> jFit <- lm(gExp ~ gType * age, jDat)
> summary(jFit)
```

```
Call:
lm(formula = gExp ~ gType * age, data = jDat)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.05383	-0.41194	-0.02491	0.31295	1.14417

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	8.07337	0.16552	48.776	< 2e-16	***
gTypeNrlKO	0.13148	0.24070	0.546	0.588	
age	-0.06818	0.01215	-5.612	2.51e-06	***
gTypeNrlKO:age	0.01019	0.01744	0.584	0.563	

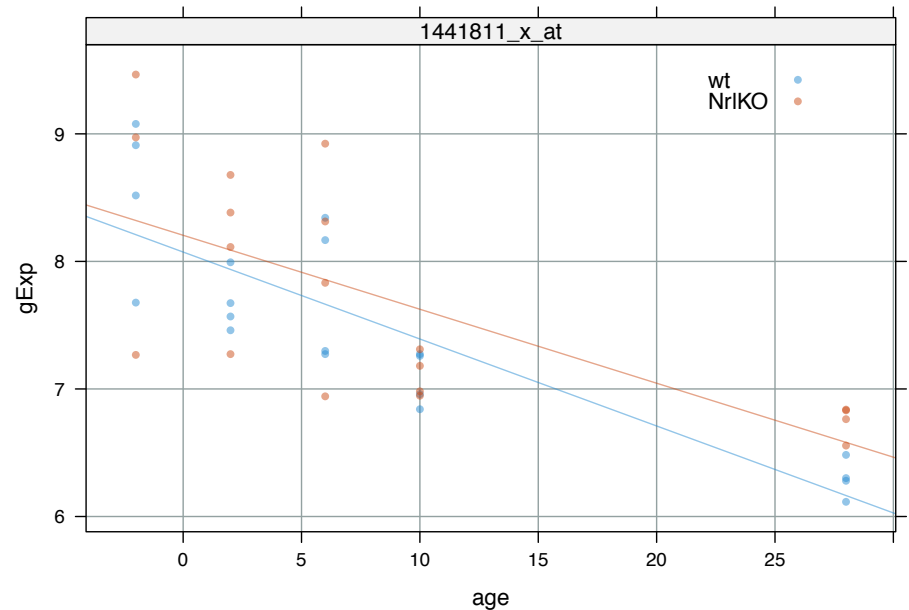
---

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

Residual standard error: 0.5651 on 35 degrees of freedom

Multiple R-squared: 0.607, Adjusted R-squared: 0.5733

F-statistic: 18.02 on 3 and 35 DF, p-value: 3.047e-07



The intercept for the knockouts is:

$$\alpha_{0,wt} + \tau_{0,\Delta Nrl}$$

and the slope for knockouts is:

$$\alpha_{1,wt} + \tau_{1,\Delta Nrl}$$

There is also a simpler way to parameterize this: but as always, you need to \*know\* how you parameterize in order to interpret your results.

$$y_{ij} = \alpha_{0,j} + \alpha_{1,j}age_i + \varepsilon_{ij}$$

where  $j \in \{wt, NrlKO\}$

$$i = 1, 2, \dots, n_j$$

```
> jFitAlt <- lm(gExp ~ gType/age - 1, jDat)
> summary(jFitAlt)
```

Call:

```
lm(formula = gExp ~ gType/age - 1, data = jDat)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.05383	-0.41194	-0.02491	0.31295	1.14417

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
gTypewt	8.07337	0.16552	48.776	< 2e-16 ***
gTypeNrlKO	8.20485	0.17476	46.949	< 2e-16 ***
gTypewt:age	-0.06818	0.01215	-5.612	2.51e-06 ***
gTypeNrlKO:age	-0.05799	0.01251	-4.636	4.80e-05 ***

---

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

Residual standard error: 0.5651 on 35 degrees of freedom

Multiple R-squared: 0.9951, Adjusted R-squared: 0.9945

F-statistic: 1761 on 4 and 35 DF, p-value: < 2.2e-16

(intercept, slope) for wild type:

$(\alpha_{0,wt}, \alpha_{1,wt})$

(intercept, slope) for the knockouts:

$(\alpha_{0,\Delta Nrl}, \alpha_{1,\Delta Nrl})$

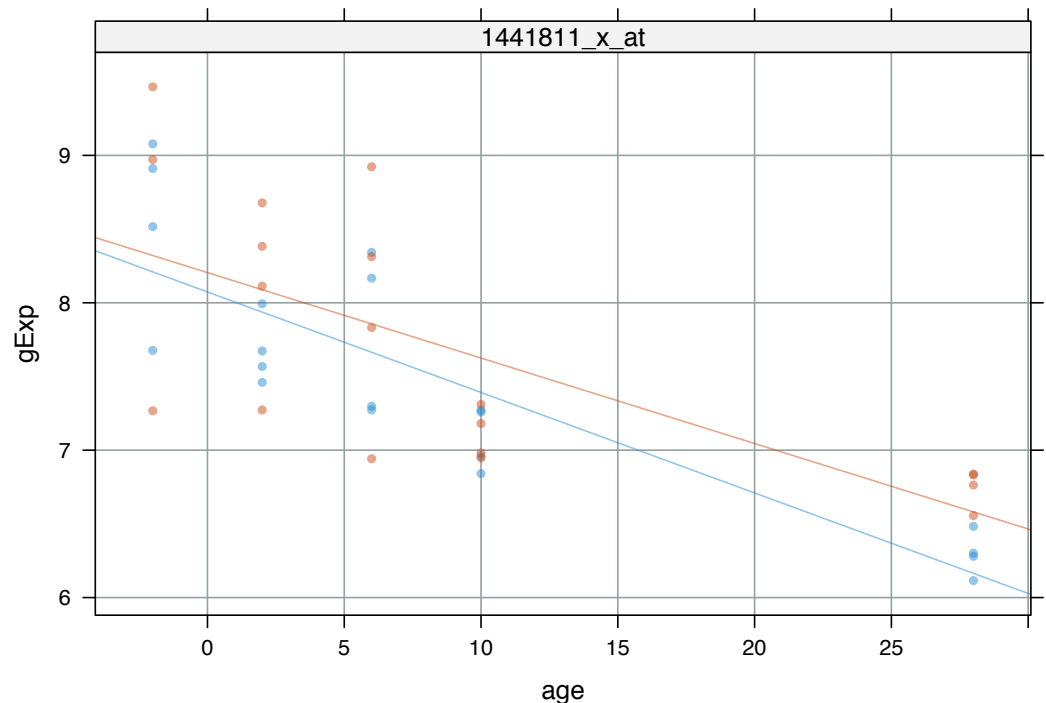


As before, you can use the F-test to assess the relevance/ effect of several terms at once.

$$F = \frac{\left( \frac{RSS_{small} - RSS_{big}}{p_{big} - p_{small}} \right)}{\frac{RSS_{big}}{n - p_{big}}} \sim_{H_0} F_{(p_{big} - p_{small}, n - p_{big})}$$

Model with and without genotype term

	small	big	
> anova(lm(gExp ~ age, jDat), jFit)			
Analysis of Variance Table			
Model 1: gExp ~ age			
Model 2: gExp ~ gType * age			
	Res.Df	RSS Df Sum of Sq F Pr(>F)	
1	37	11.774	
2	35	11.176	2 0.59807 0.9365 0.4016



Let's go back and consider a more complex relationship between gene expression and age

- The nature of the regression function  $f(x; \alpha)$  is one of the defining characteristics of a regression model
  - $f$  linear in  $\alpha \Rightarrow$  linear model
  - $f$  not linear in  $\alpha \Rightarrow$  nonlinear model

nonlinear parametric regression

$$Y = \frac{1}{1 + e^{(\phi - x)/\xi}} + \varepsilon$$

simple linear regression (a linear model)

$$Y = \alpha_0 + \alpha_1 x + \varepsilon$$

What we just did.



polynomial regression (also a linear model)

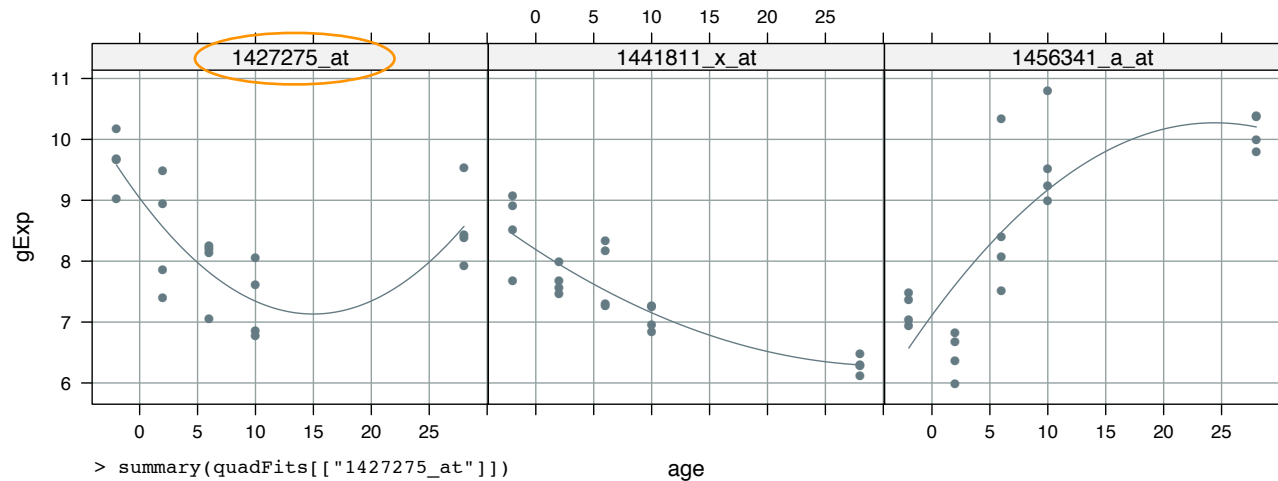
$$Y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \varepsilon$$

What we're  
about to do.



Recall the polynomial function of degree \*n\*. We will focus on cubic polynomial (degree 2)

$$Y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \varepsilon$$



Call:

```
lm(formula = gExp ~ age + I(age^2), data = z)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.16275	-0.55506	0.09503	0.40804	0.95803

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	9.036401	0.212313	42.562	< 2e-16 ***
age	-0.254305	0.048125	-5.284	6.07e-05 ***
I(age^2)	0.008490	0.001661	5.110	8.71e-05 ***

---

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

Residual standard error: 0.6444 on 17 degrees of freedom

Multiple R-squared: 0.6218, Adjusted R-squared: 0.5773

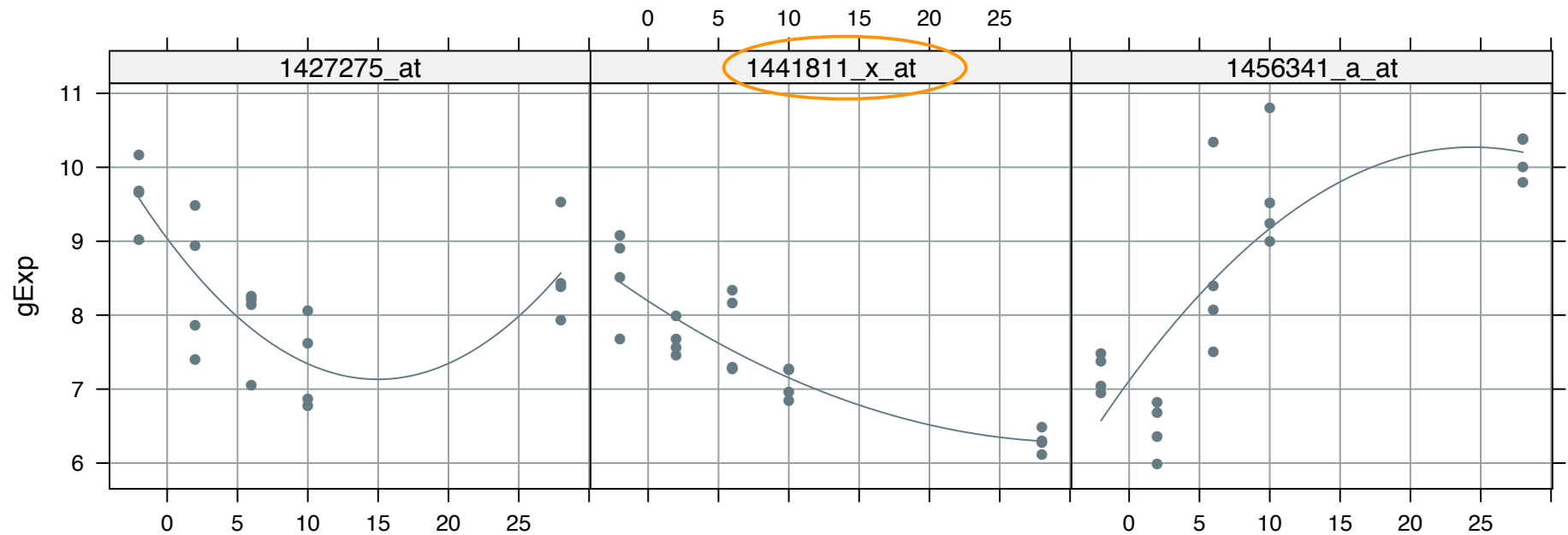
F-statistic: 13.98 on 2 and 17 DF, p-value: 0.0002572

- The nature of the regression function  $f(x; \alpha)$  is one of the defining characteristics of a regression model
  - $f$  linear in  $\alpha \Rightarrow$  linear model
  - $f$  not linear in  $\alpha \Rightarrow$  nonlinear model

polynomial regression (also a linear model)

$$Y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \varepsilon$$

**NOTE: This is a linear model**, because it is linear in the alphas. It is easy but wrong to focus on the  $x$ 's and mistake this for a nonlinear model.



```
> summary(quadFits[["1441811_x_at"]])
```

age

Call:

```
lm(formula = gExp ~ age + I(age^2), data = z)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.76946	-0.25477	-0.00589	0.13662	0.82202

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	8.190766	0.140969	58.103	< 2e-16 ***
age	-0.123836	0.031953	-3.876	0.00121 **
I(age^2)	0.002006	0.001103	1.819	0.08660 .

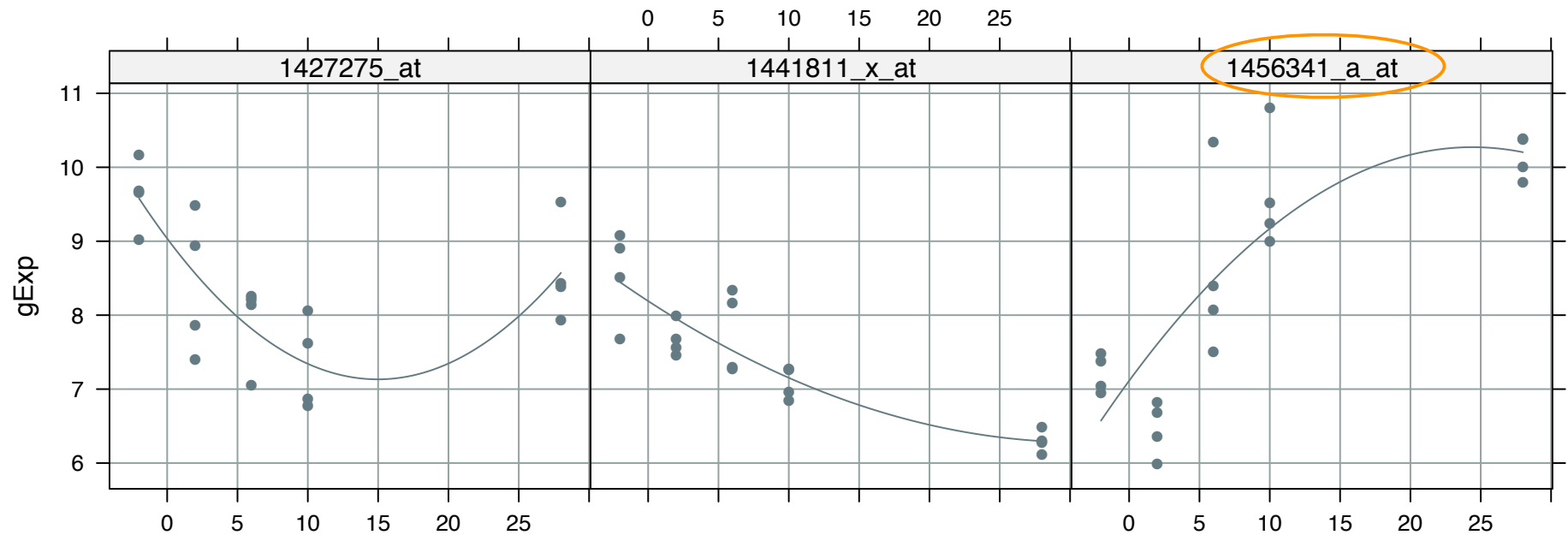
---

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

Residual standard error: 0.4279 on 17 degrees of freedom

Multiple R-squared: 0.774, Adjusted R-squared: 0.7475

F-statistic: 29.12 on 2 and 17 DF, p-value: 3.23e-06



```
> summary(quadFits[["1456341_a_at"]])
```

age

Call:

```
lm(formula = gExp ~ age + I(age^2), data = z)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.6211	-0.5010	-0.0050	0.3955	1.8651

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.112481	0.310922	22.875	3.3e-14 ***
age	0.258892	0.070477	3.673	0.00188 **
I(age^2)	-0.005303	0.002433	-2.180	0.04363 *

---

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

Residual standard error: 0.9437 on 17 degrees of freedom

Multiple R-squared: 0.6737, Adjusted R-squared: 0.6353

F-statistic: 17.55 on 2 and 17 DF, p-value: 7.337e-05

# How can we tell if polynomial is \*better\* than simple linear?

## F tests in regression

Remember this?

small model is nested within big, e.g., it's a special case where some parameters are equal to zero

model	example	# params = DF	RSS
small	$\text{lm}(y \sim \text{gType} + \text{devStage})$	$p_{\text{small}} = 6$	$\text{RSS}_{\text{small}}$
big	$\text{lm}(y \sim \text{gType} * \text{devStage})$	$p_{\text{big}} = 10$	$\text{RSS}_{\text{big}}$

$$y_{ijk} = \theta + \tau_j + \beta_k + (\tau\beta)_{jk} + \varepsilon_{ijk} \text{ "big"}$$

$$y_{ijk} = \theta + \tau_j + \beta_k + (\tau\beta)_{jk} + \varepsilon_{ijk} \text{ "small"}$$

by definition:

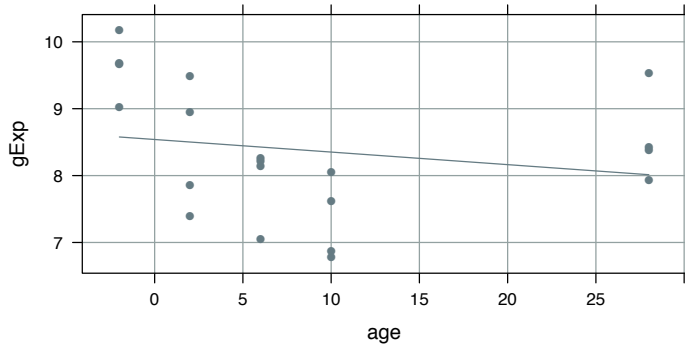
$$p_{\text{small}} < p_{\text{big}}$$

$$\text{RSS}_{\text{small}} \geq \text{RSS}_{\text{big}}$$

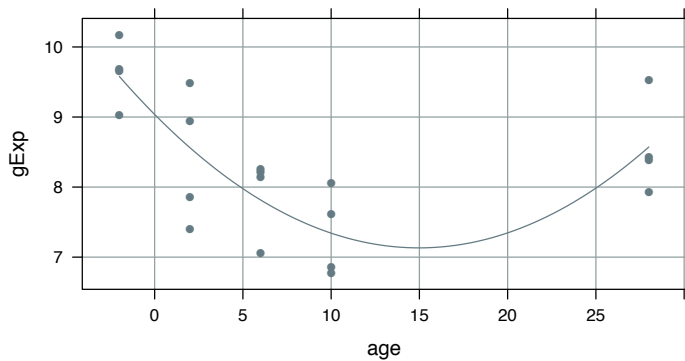
$$F = \frac{\left( \frac{\text{RSS}_{\text{small}} - \text{RSS}_{\text{big}}}{p_{\text{big}} - p_{\text{small}}} \right)}{\frac{\text{RSS}_{\text{big}}}{n - p_{\text{big}}}} \sim_{H_0} F_{(p_{\text{big}} - p_{\text{small}}, n - p_{\text{big}})}$$



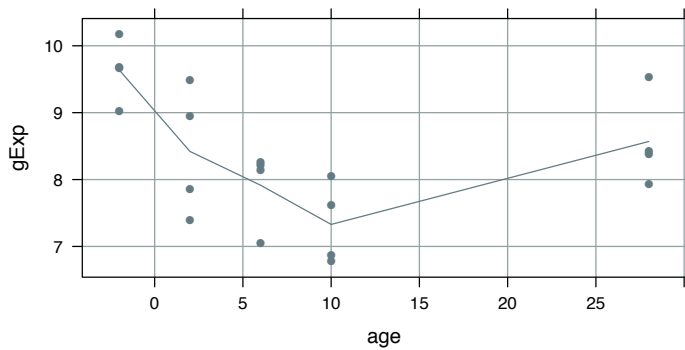
How many parameters do each of these models have?



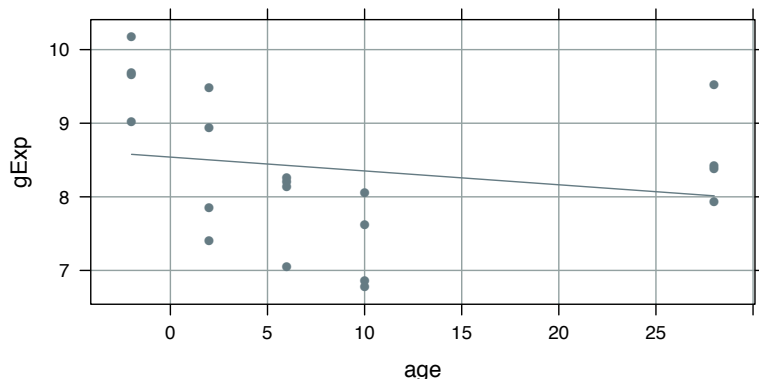
Linear



Polynomial regression



Anova: categorical variable



```
> (jGene <- luckyGenes[1])
```

```
[1] "1427275_at"
```

**small**

**big**

```
> anova(linFits[[jGene]], quadFits[[jGene]])
```

Analysis of Variance Table

Model 1:  $gExp \sim age$

Model 2:  $gExp \sim age + I(age^2)$

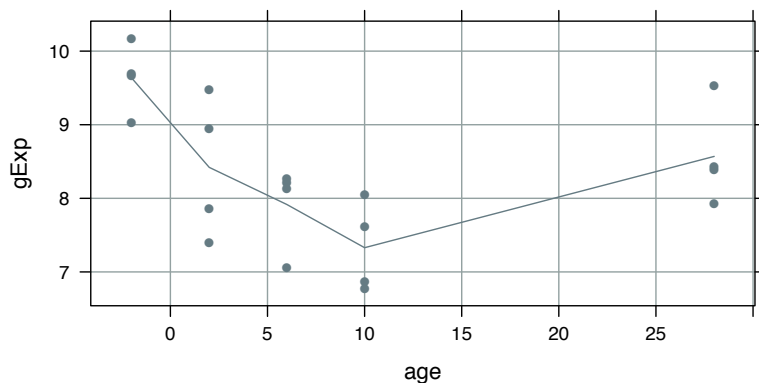
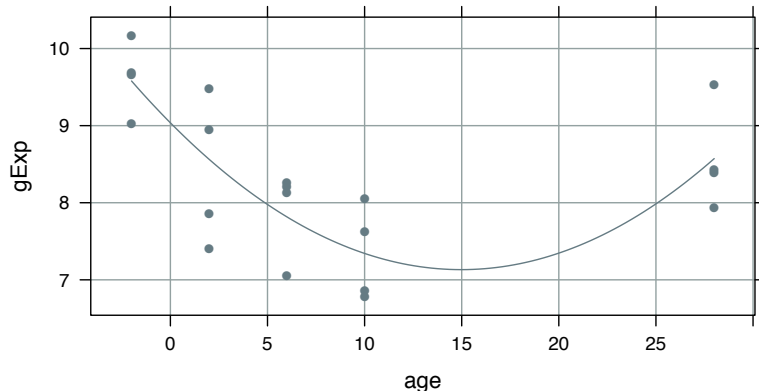
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	18	17.9021				
2	17	7.0591	1	10.843	26.113	8.71e-05 ***

---

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

```
> AIC(linFits[[jGene]], quadFits[[jGene]], factFits[[jGene]])
```

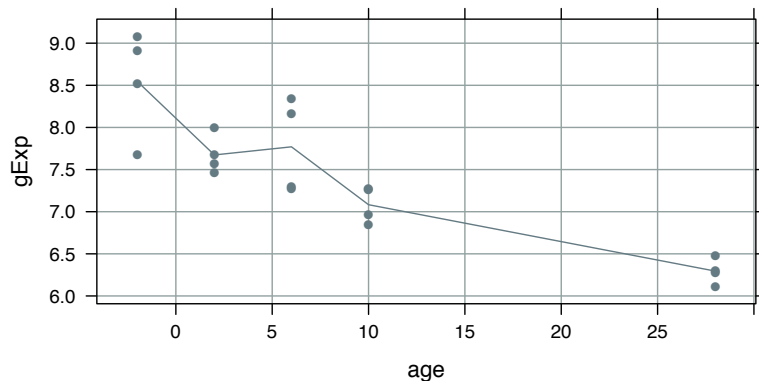
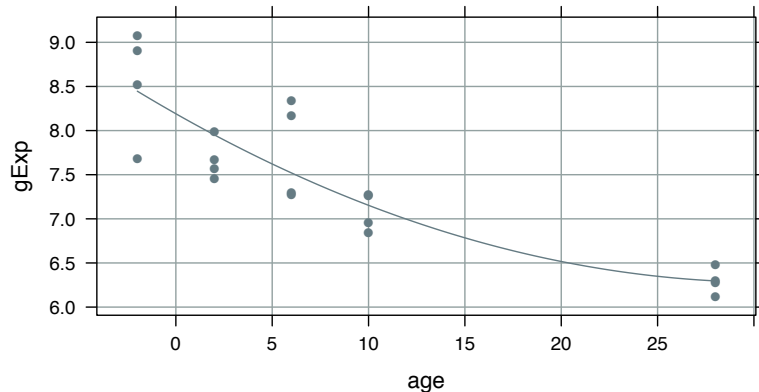
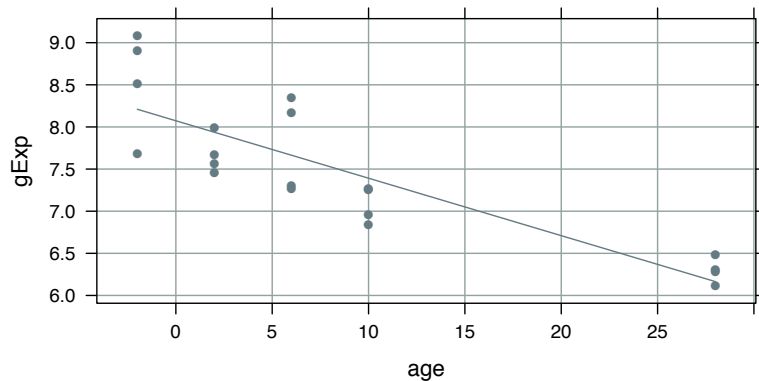
	df	AIC
linFits[[jGene]]	3	60.54129
quadFits[[jGene]]	4	43.92930
factFits[[jGene]]	6	47.54810



it's “worth it” to go from linear to quadratic here

but hard to justify going from quadratic to one-way ANOVA

possible links to read more about using AIC to compare non-nested models: [stackexchange](#) and [Wikipedia](#)



```
> (jGene <- luckyGenes[3])
[1] "1441811_x_at"
```

small

big

```
> anova(linFits[[jGene]], quadFits[[jGene]])
```

Analysis of Variance Table

Model 1: gExp ~ age

Model 2: gExp ~ age + I(age^2)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	18	3.7176				
2	17	3.1120	1	0.60559	3.3081	0.0866 .

---

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

```
> AIC(linFits[[jGene]], quadFits[[jGene]], factFits[[jGene]])
```

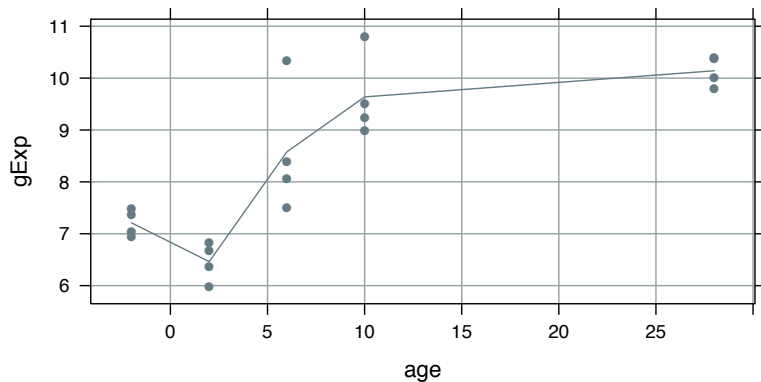
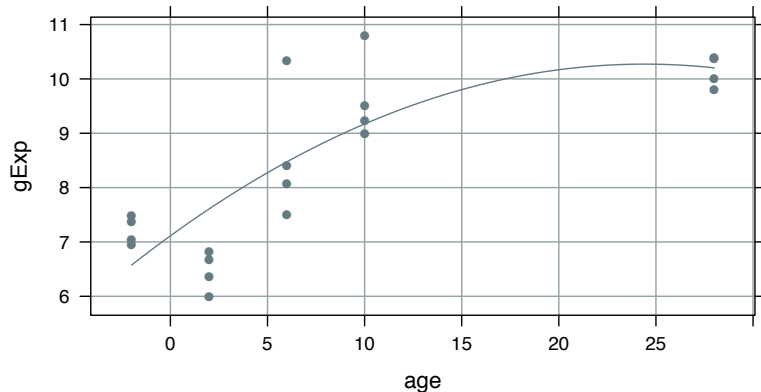
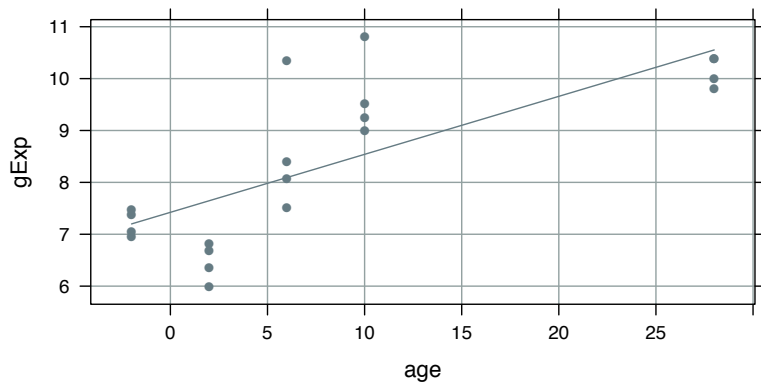
	df	AIC
linFits[[jGene]]	3	29.10466
quadFits[[jGene]]	4	27.54851
factFits[[jGene]]	6	27.12587

meh

not clear it's “worth it” to go from linear to quadratic here

even less payoff to go from quadratic to one-way ANOVA

Occam's Razor and the KISS principle → stick w/ simple linear model



```
> (jGene <- luckyGenes[2])
```

```
[1] "1456341_a_at"
```

small

big

```
> anova(linFits[[jGene]], quadFits[[jGene]])
```

Analysis of Variance Table

Model 1: gExp ~ age

Model 2: gExp ~ age + I(age^2)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	18	19.370				
2	17	15.139	1	4.2308	4.7509	0.04363 *

---

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

```
> AIC(linFits[[jGene]], quadFits[[jGene]], factFits[[jGene]])
```

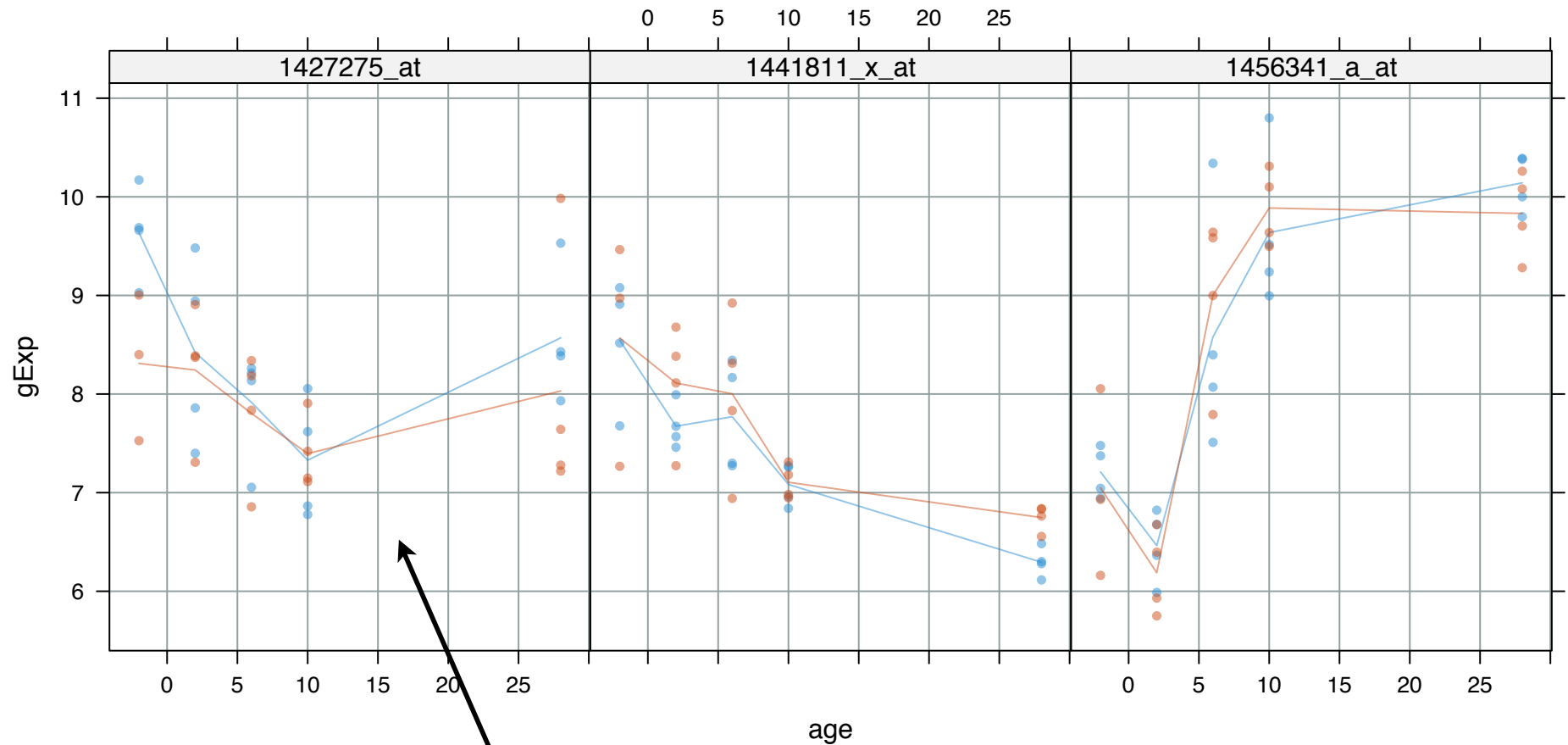
	df	AIC
linFits[[jGene]]	3	62.11743
quadFits[[jGene]]	4	59.18864
factFits[[jGene]]	6	48.70210

it's probably "worth it" to go from linear to quadratic here (?)

going from quadratic to one-way ANOVA seems justified

Let's make it more complex: a quadratic model  
which includes genotype effect

i.e., one categorical and continuous variable,  
with continuous variable modeled quadratically



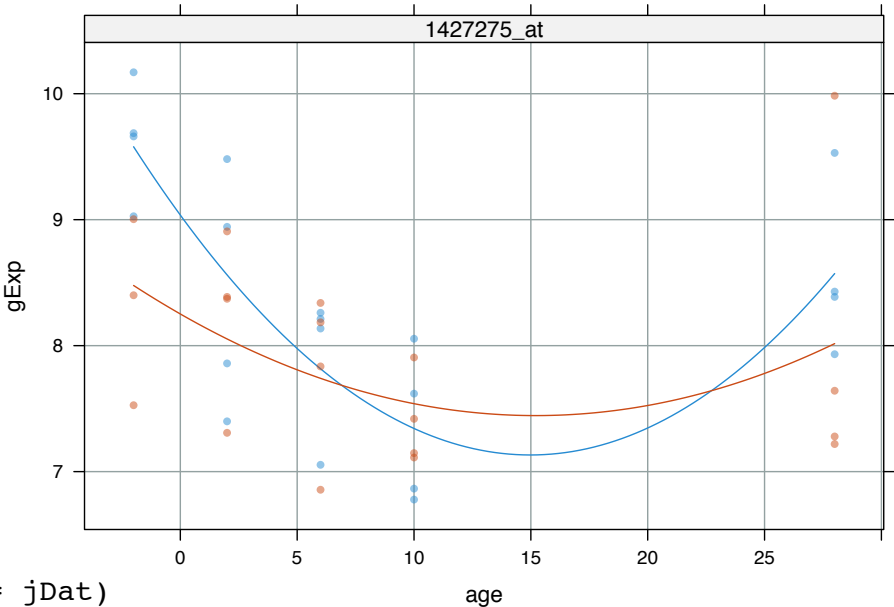
let's focus here for a model including a quadratic age term

$$y_{ij} = \alpha_{0,wt} + \tau_{0,j} + (\alpha_{1,wt} + \tau_{1,j})age_i + (\alpha_{2,wt} + \tau_{2,j})age_i^2 + \varepsilon_{ij}$$

where  $j \in \{wt, NrlKO\}$

$i = 1, 2, \dots, n_j$

$$\tau_{0,wt} = \tau_{1,wt} = \tau_{2,wt} \equiv 0$$



```
> summary(jFit)
```

```
Call:
lm(formula = gExp ~ gType * (age + I(age^2)), data = jDat)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.16275	-0.55816	0.08203	0.42020	1.96803

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	9.036401	0.234853	38.477	< 2e-16	***
gTypeNrlKO	-0.784969	0.350249	-2.241	0.0319	*
age	-0.254305	0.053234	-4.777	3.55e-05	***
I(age^2)	0.008490	0.001838	4.620	5.63e-05	***
gTypeNrlKO:age	0.148195	0.078232	1.894	0.0670	.
gTypeNrlKO:I(age^2)	-0.005001	0.002673	-1.871	0.0702	.

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

Residual standard error: 0.7128 on 33 degrees of freedom  
Multiple R-squared: 0.4755, Adjusted R-squared: 0.3961  
F-statistic: 5.984 on 5 and 33 DF, p-value: 0.0004804

as always, you can assess the relevance of several terms at once -- such as everything involving genotype -- with an F test

borderline evidence that genotype affects something about the parabola (location or shape)

small

```
> anova(lm(gExp ~ age + I(age^2), jDat),  
+       lm(gExp ~ gType * (age + I(age^2)), jDat))
```

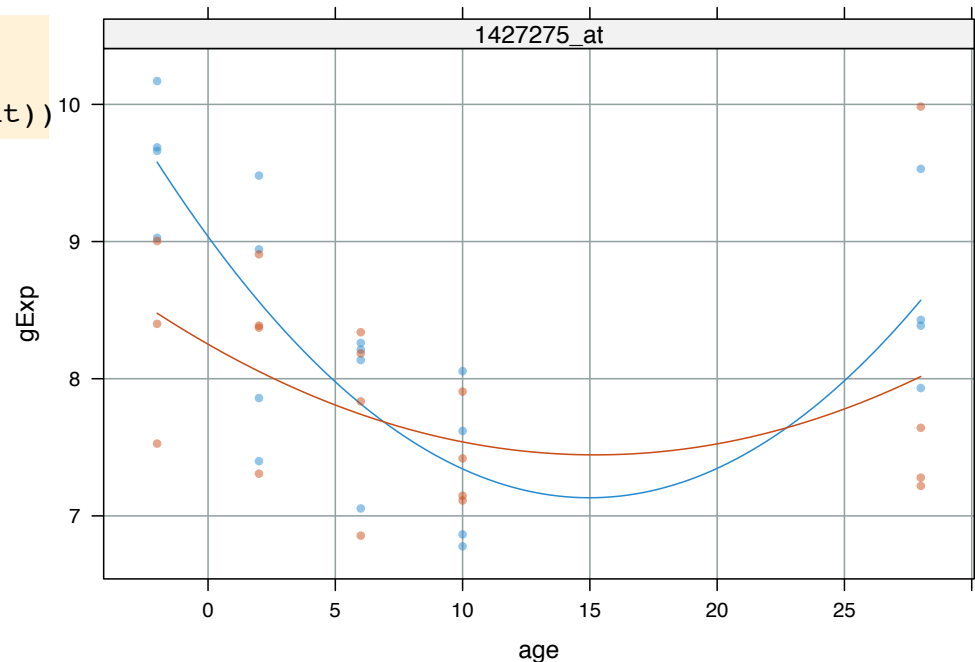
big

Analysis of Variance Table

Model 1:  $\text{gExp} \sim \text{age} + \text{I}(\text{age}^2)$

Model 2:  $\text{gExp} \sim \text{gType} * (\text{age} + \text{I}(\text{age}^2))$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	36	20.081				
2	33	16.767	3	3.3144	2.1744	0.1097





linear model framework is extremely general!

one extreme (simple): two-sample common variance t-test

another extreme (flexible): a polynomial, potentially different for each level of some factor

dichotomous variable? OK!

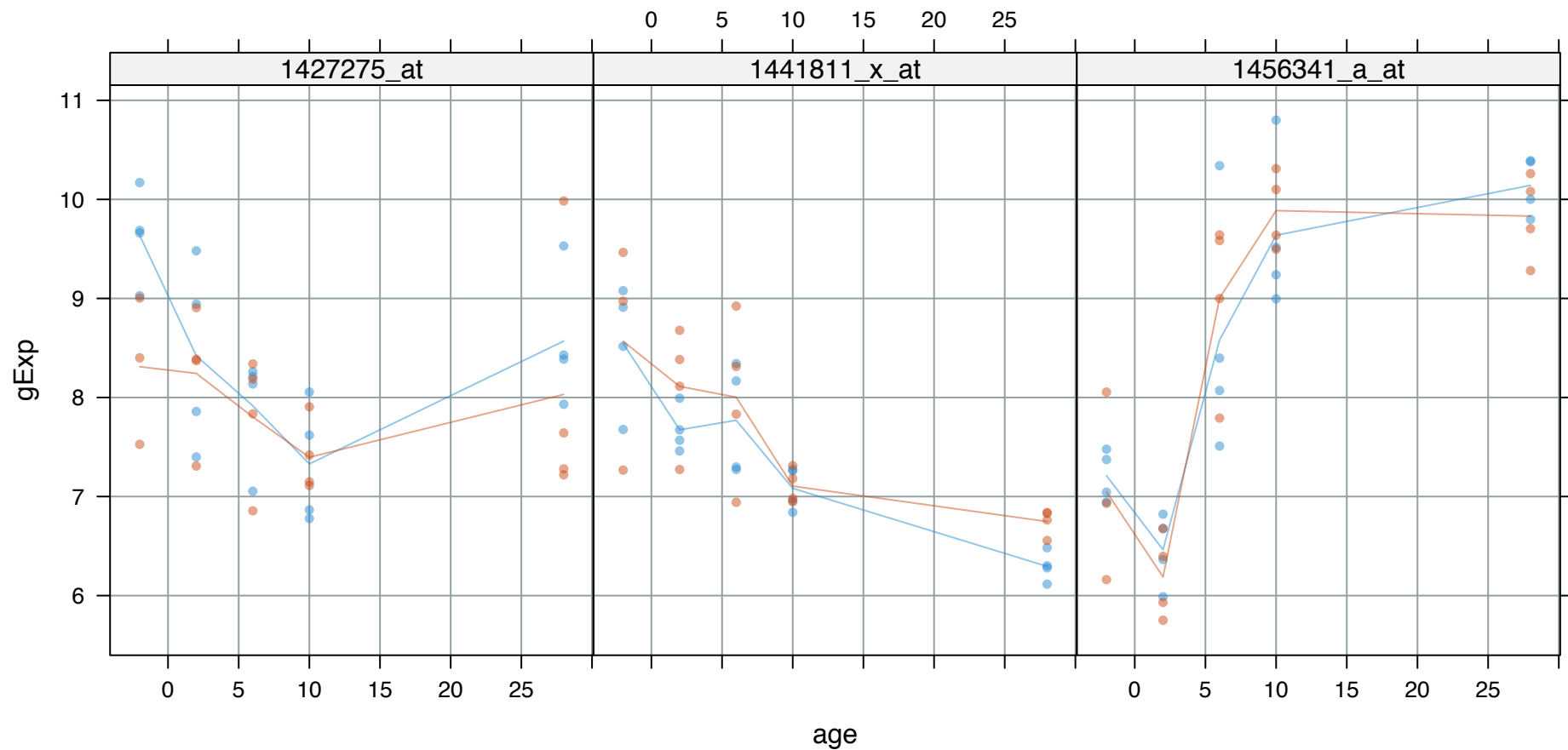
categorical variable? OK!

quantitative variable? OK!

various combinations of the above? OK!

don't be afraid to build models with more than 1 covariate

don't be intimidated by all the “contrast” talk



What about the other 29,946 probesets?

`lm(yMat ~ x)`

$$Y = X\alpha + \varepsilon$$

$$\begin{bmatrix} y_{11} & \cdots & y_{1G} \\ y_{21} & & y_{2G} \\ \vdots & & \\ y_{n1} & & y_{nG} \end{bmatrix} = X \begin{bmatrix} \alpha_1 & \cdots & \alpha_G \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1G} \\ \varepsilon_{21} & & \varepsilon_{2G} \\ \vdots & & \\ \varepsilon_{n1} & & \varepsilon_{nG} \end{bmatrix}$$

built-in function `lm()` can do “multivariate regression” = many dependent vars (“responses”)  
aka “multivariate multiple regression”

From `lm()` documentation:

If response is a matrix a linear model is fitted separately by least-squares to each column of the matrix.

`lm` returns an object of class “lm” or for multiple responses of class `c("mlm", "lm")`.

Industrial scale model fitting is good because things like this are not recomputed 30K times unnecessarily\*

$Y = X\alpha + \varepsilon$  regression model

$\hat{\alpha} = (X^T X)^{-1} X^T Y$  the MLE and OLS estimator of  $\alpha$

$\hat{\sigma}^2 = \frac{1}{n-p} \hat{\varepsilon}^T \hat{\varepsilon}$  the estimated error variance

$\hat{V}(\hat{\alpha}) = \hat{\sigma}^2 (X^T X)^{-1}$  the estimated covariance matrix of  $\hat{\alpha}$

How test  $H_0 : \alpha_j = 0$ ?

With a t-statistic. Under  $H_0$ , we have (at least approximately) that:

$$\frac{\hat{\alpha}_j}{\widehat{se}(\hat{\alpha}_j)} \sim t_{n-p}$$

so a p-value is obtained by computing a tail probability for the observed value of  $\hat{\alpha}_j$  from a  $t_{n-p}$  distribution.

\* under the hood, `lm()` is doing something more clever and numerically stable than this

I have fit all the models we've considered to all ~30K probesets.

Let's examine some of the results *en masse*.

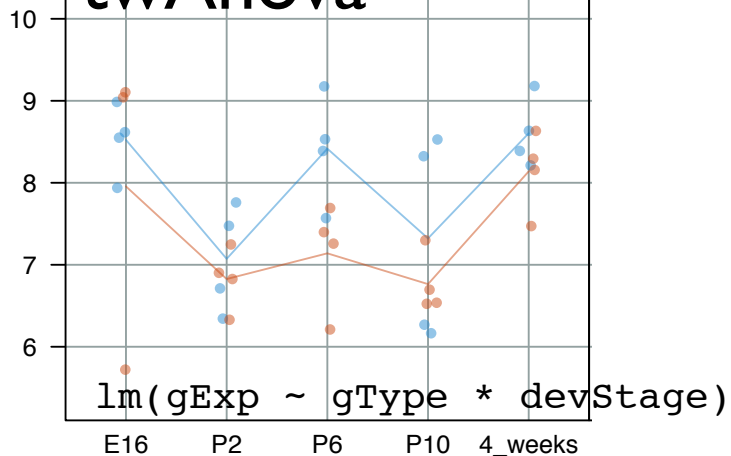
Let this drive home the point that ...

- background variability
- intercepts
- Nrl knockout effects
- devStage effects
- age effects, both linear and quadratic
- and interactions of all the above

differ for each gene.

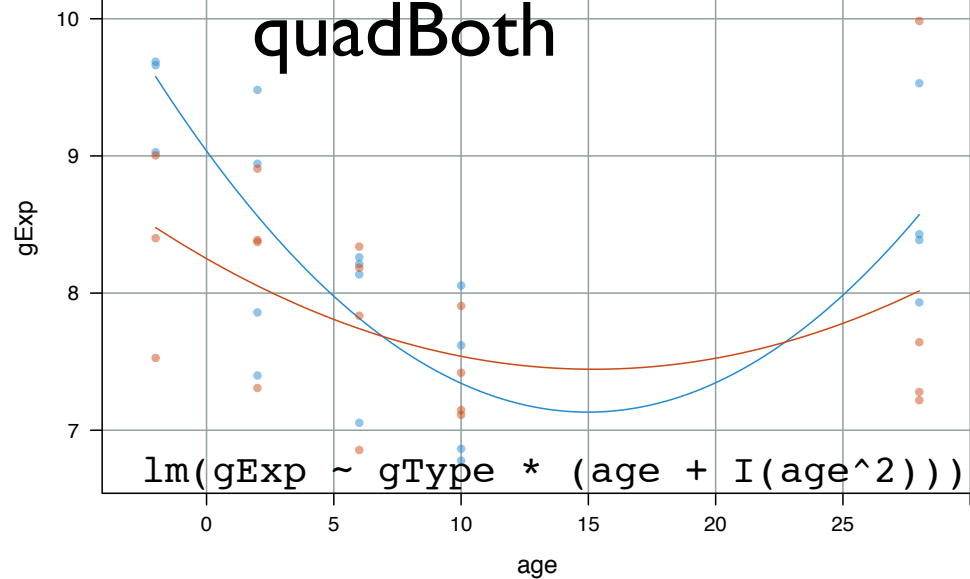
1438786\_a\_at

twAnova



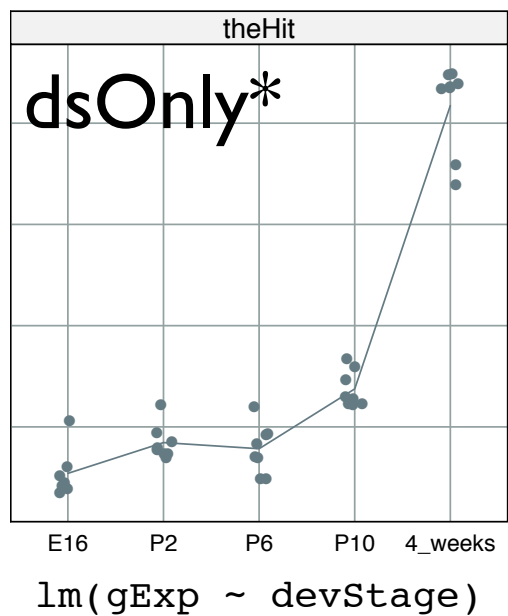
1427275\_at

quadBoth



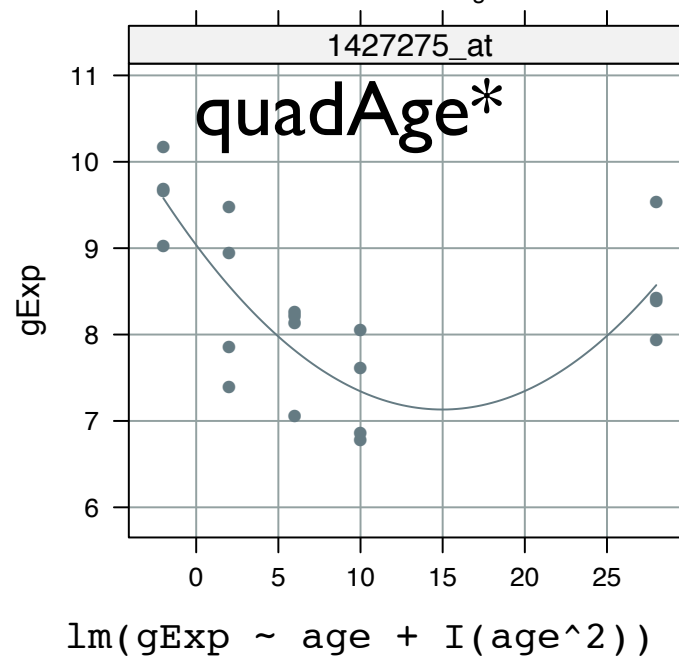
theHit

dsOnly\*



1427275\_at

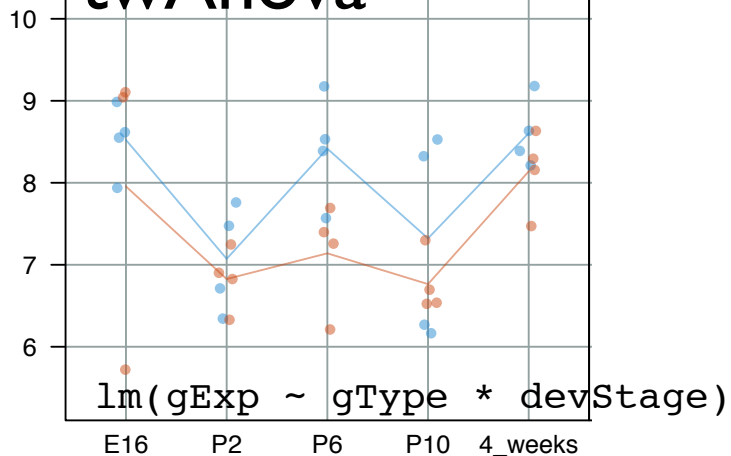
quadAge\*



\* Figures slightly misleading. Model is fit to all the data, wild type and Nrl knockout, but gType is not used as a covariate.

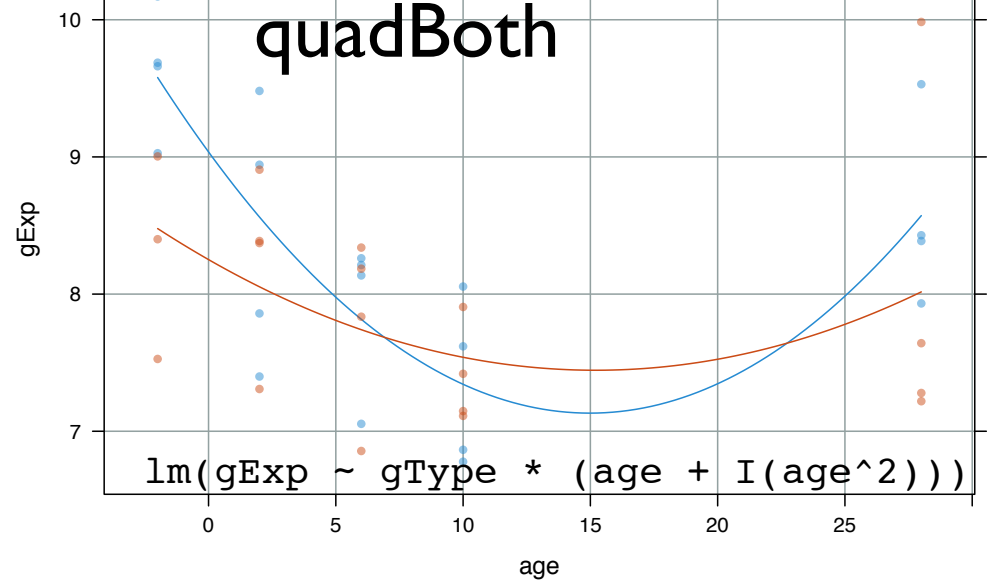
1438786\_a\_at

twAnova



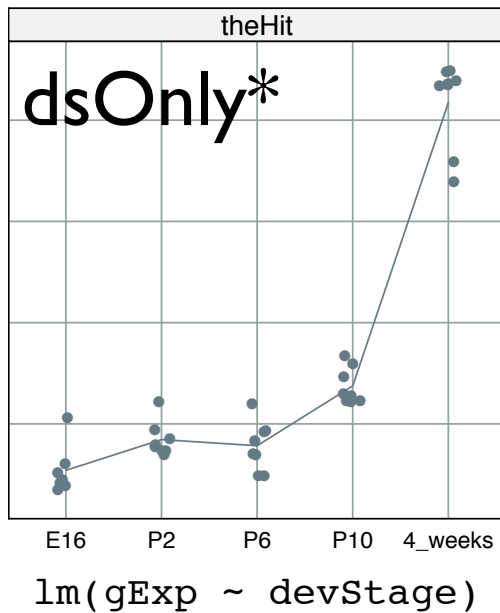
1427275\_at

quadBoth



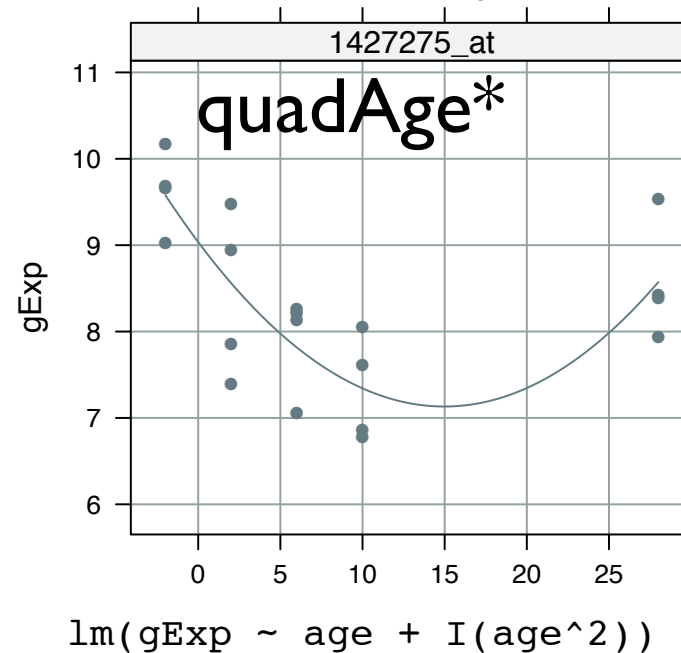
theHit

dsOnly\*



1427275\_at

quadAge\*



How “big” are these models? How many parameters are we using to specify the mean structure?

1438786\_a\_at

twAnova

10

lm(gExp ~ gType \* devStage)

E16 P2 P6 P10 4\_weeks

1427275\_at

quadBoth

6

lm(gExp ~ gType \* (age + I(age^2)))

gExp

age

theHit

dsOnly\*

5

lm(gExp ~ devStage)

E16 P2 P6 P10 4\_weeks

1427275\_at

quadAge\*

3

lm(gExp ~ age + I(age^2))

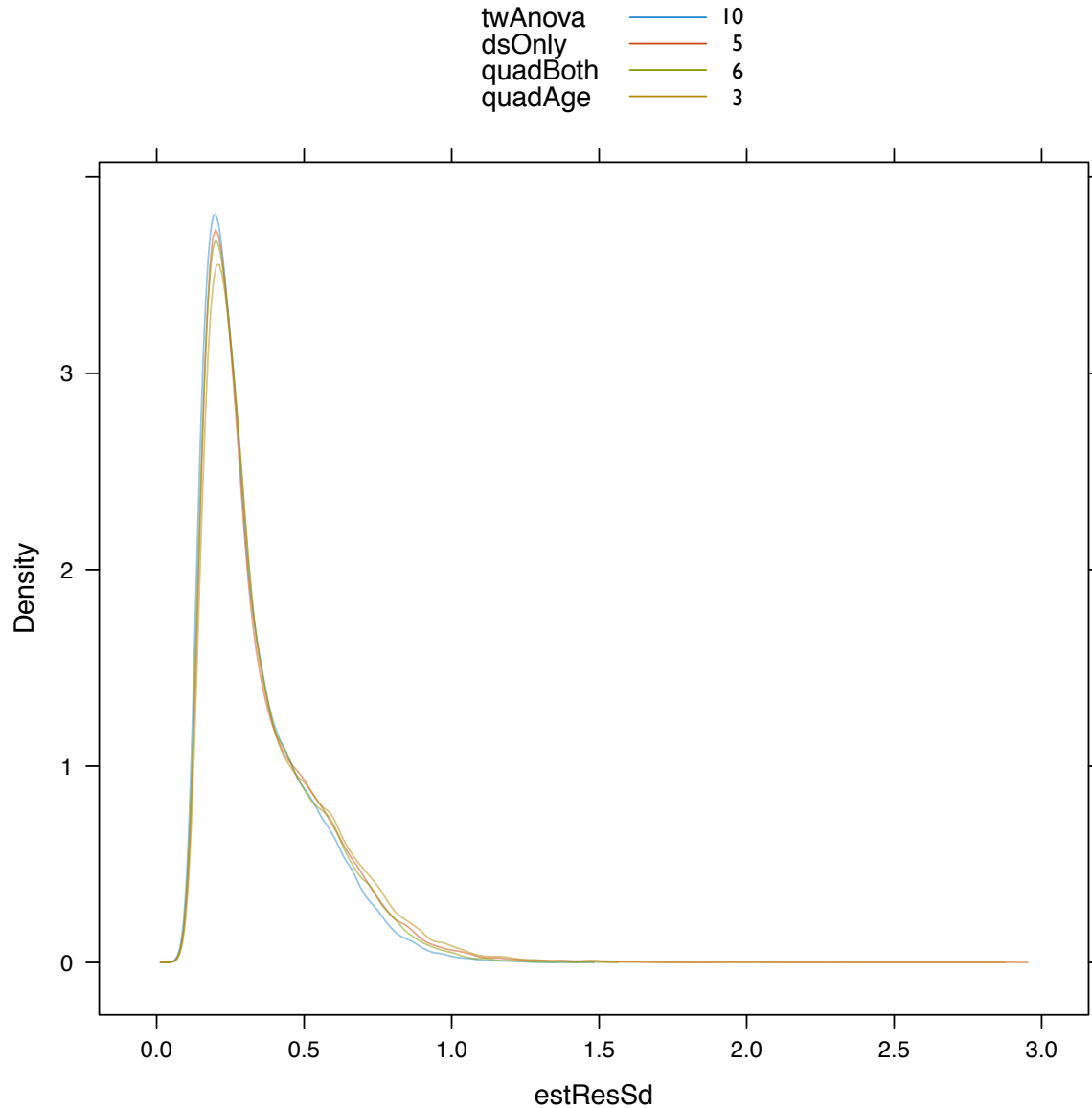
gExp

How “big” are these models? How many parameters are we using to specify the mean structure?



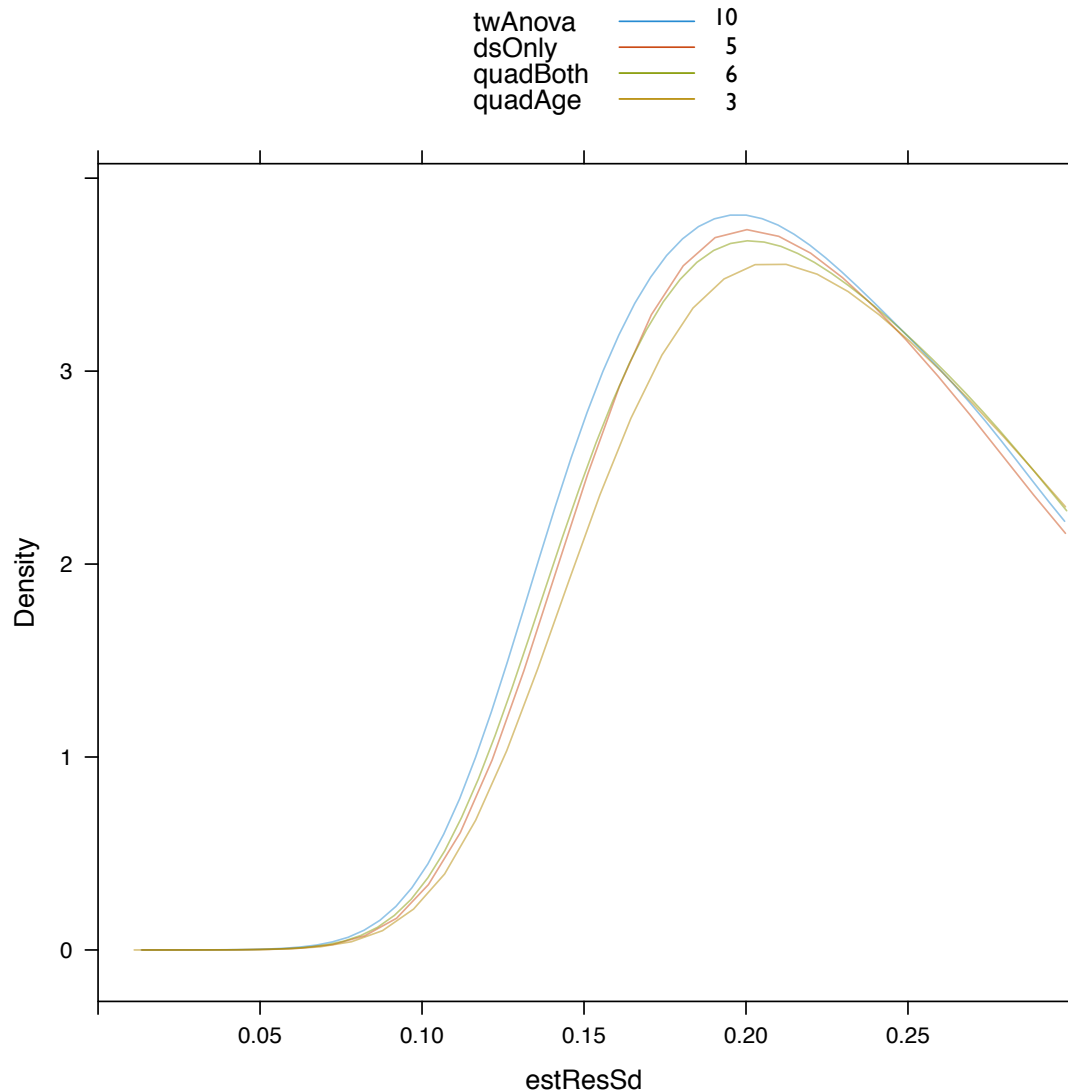
$$y_i = f(x_i; \alpha) + \varepsilon_i, \text{var}(\varepsilon) = \sigma^2$$

Let's look at estimates of the error standard deviation.



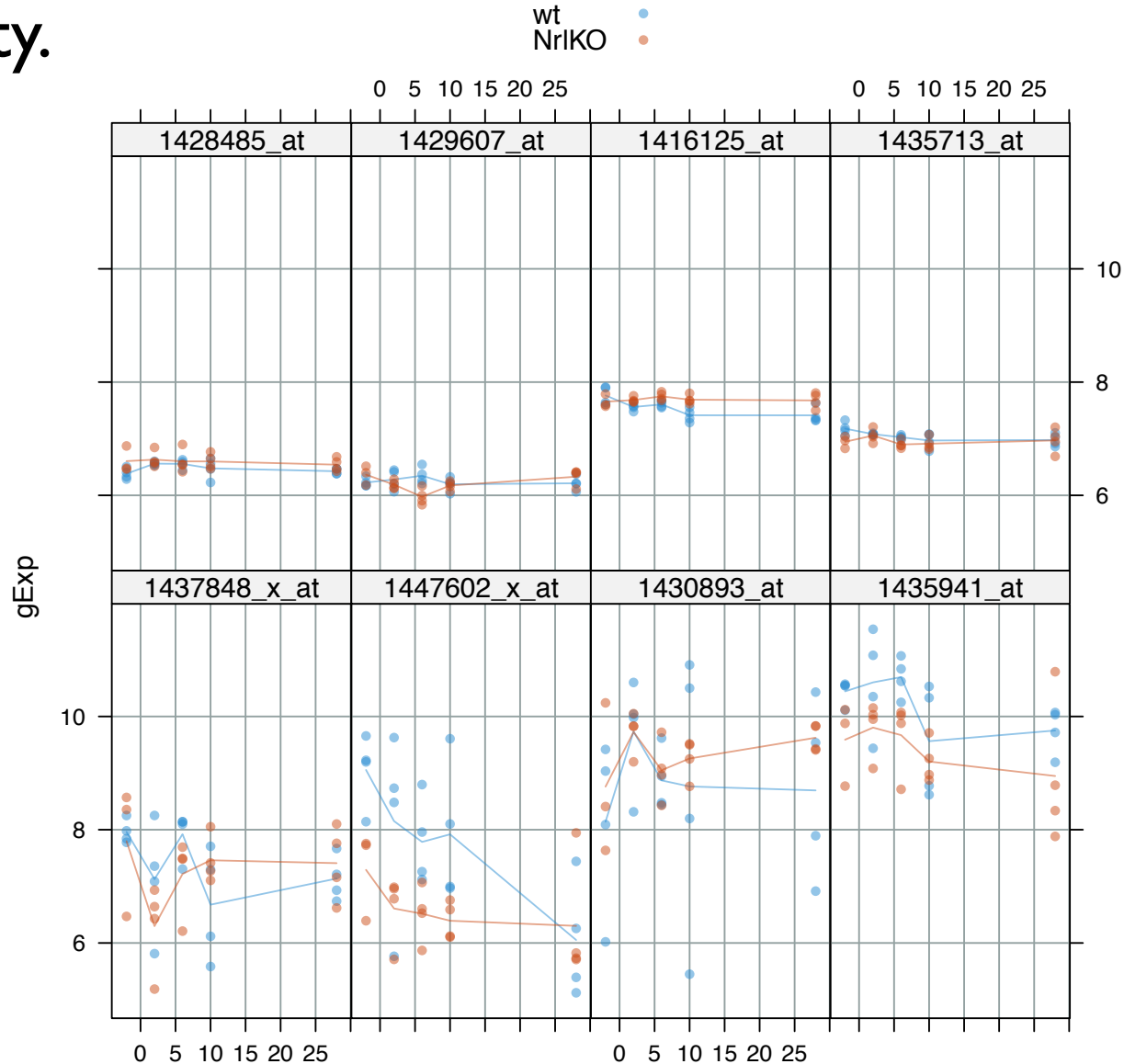
$$y_i = f(x_i; \alpha) + \varepsilon_i, \text{var}(\varepsilon) = \sigma^2$$

Let's look at estimates of the error standard deviation.



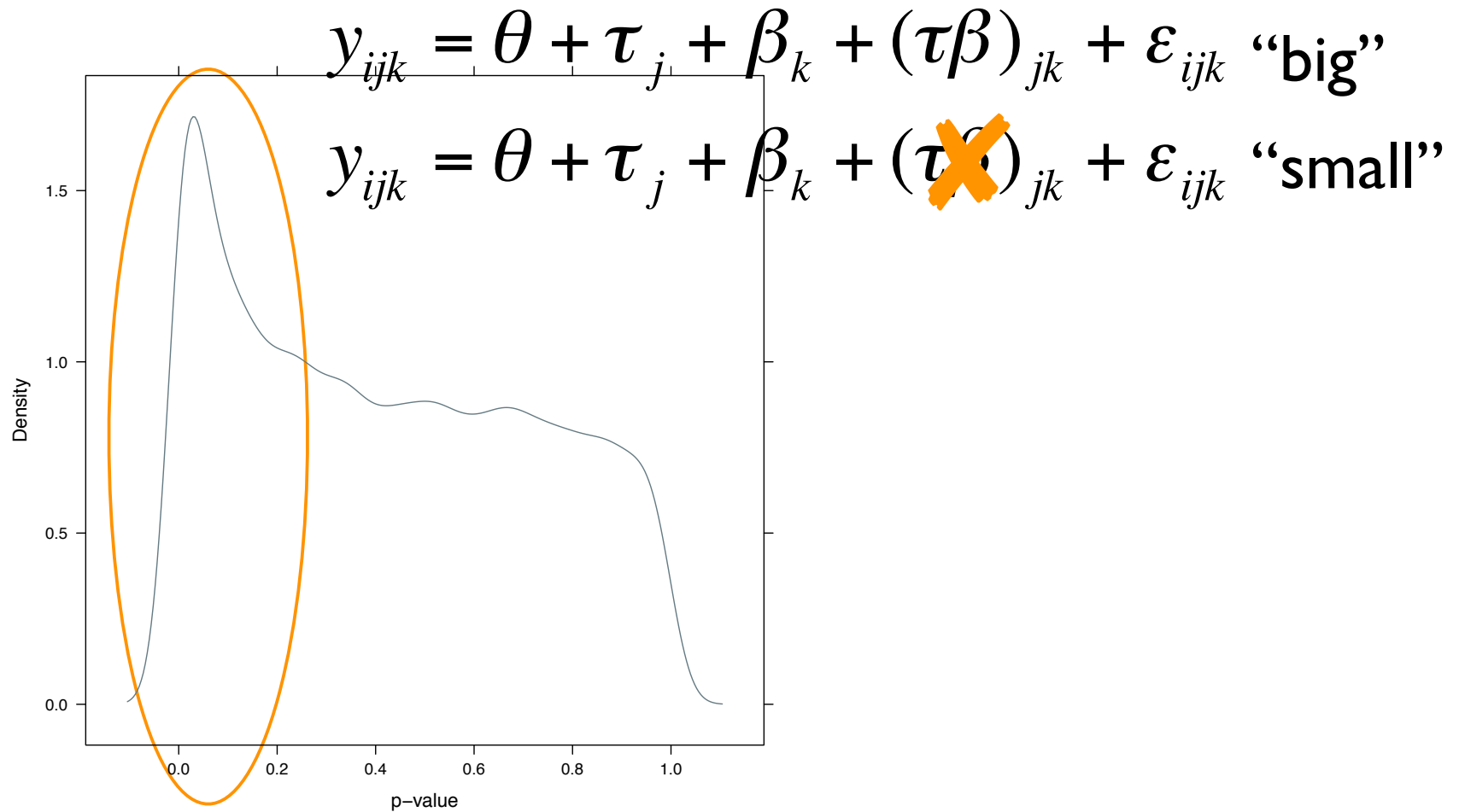
$$y_i = f(x_i; \alpha) + \varepsilon_i, \text{var}(\varepsilon) = \sigma^2$$

Let's look genes exhibiting extremely low or high variability.



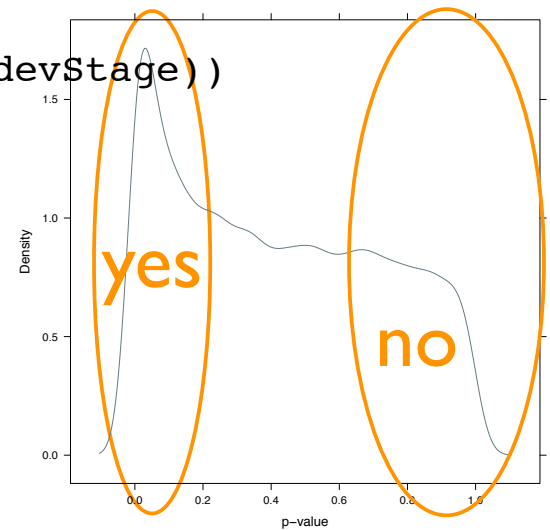
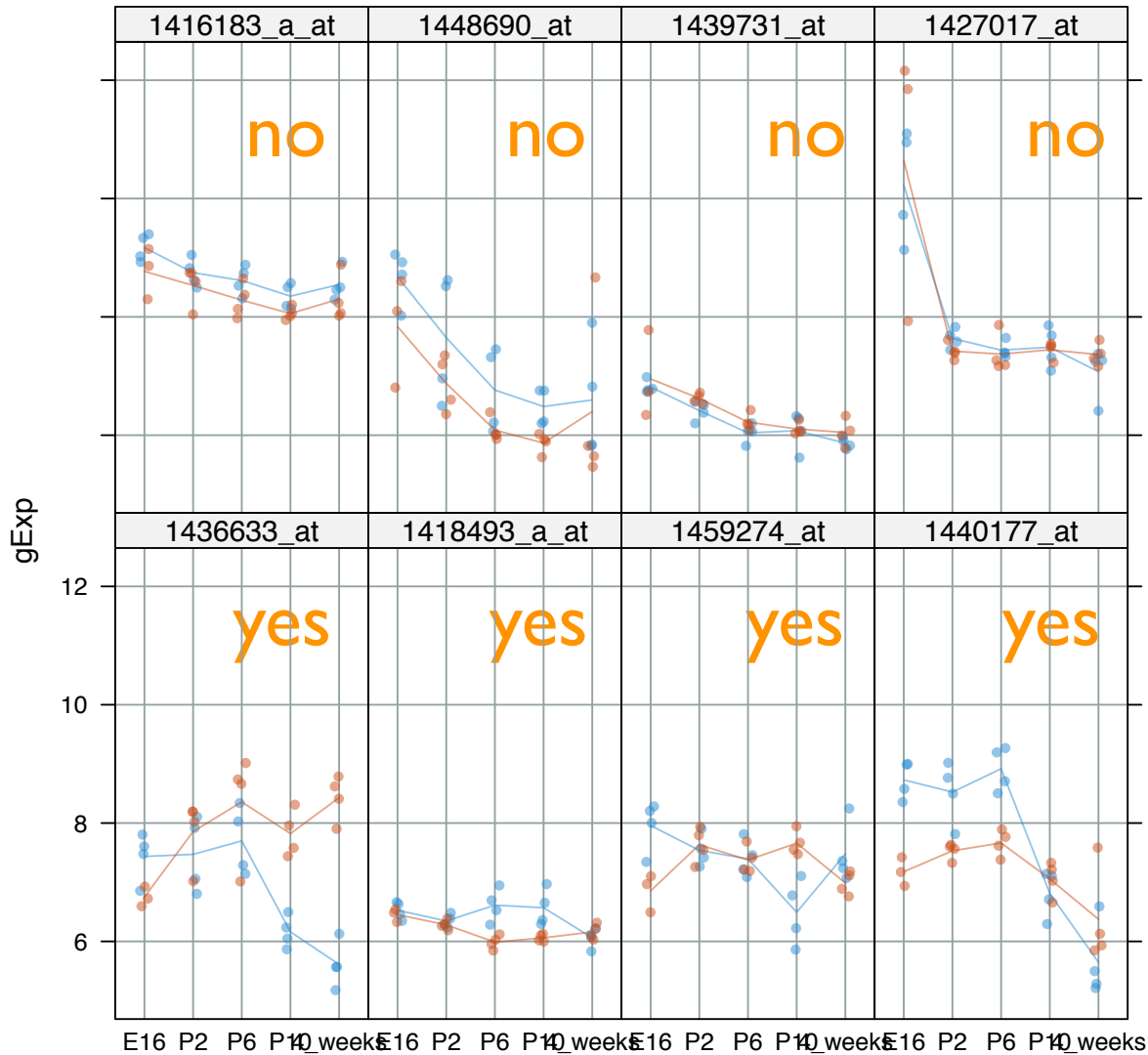
# In the two-way ANOVA model, is there evidence for gType \* devStage interaction? YES.

```
## this code is fictional but conveys the point  
anova(lm(gExp ~ gType * devStage), lm(gExp ~ gType + devStage))  
## inspecting the p-values from these F tests
```



```
## this code is fictional but conveys the point
anova(lm(gExp ~ gType * devStage), lm(gExp ~ gType + devStage))
## inspecting the p-values from these F tests
```

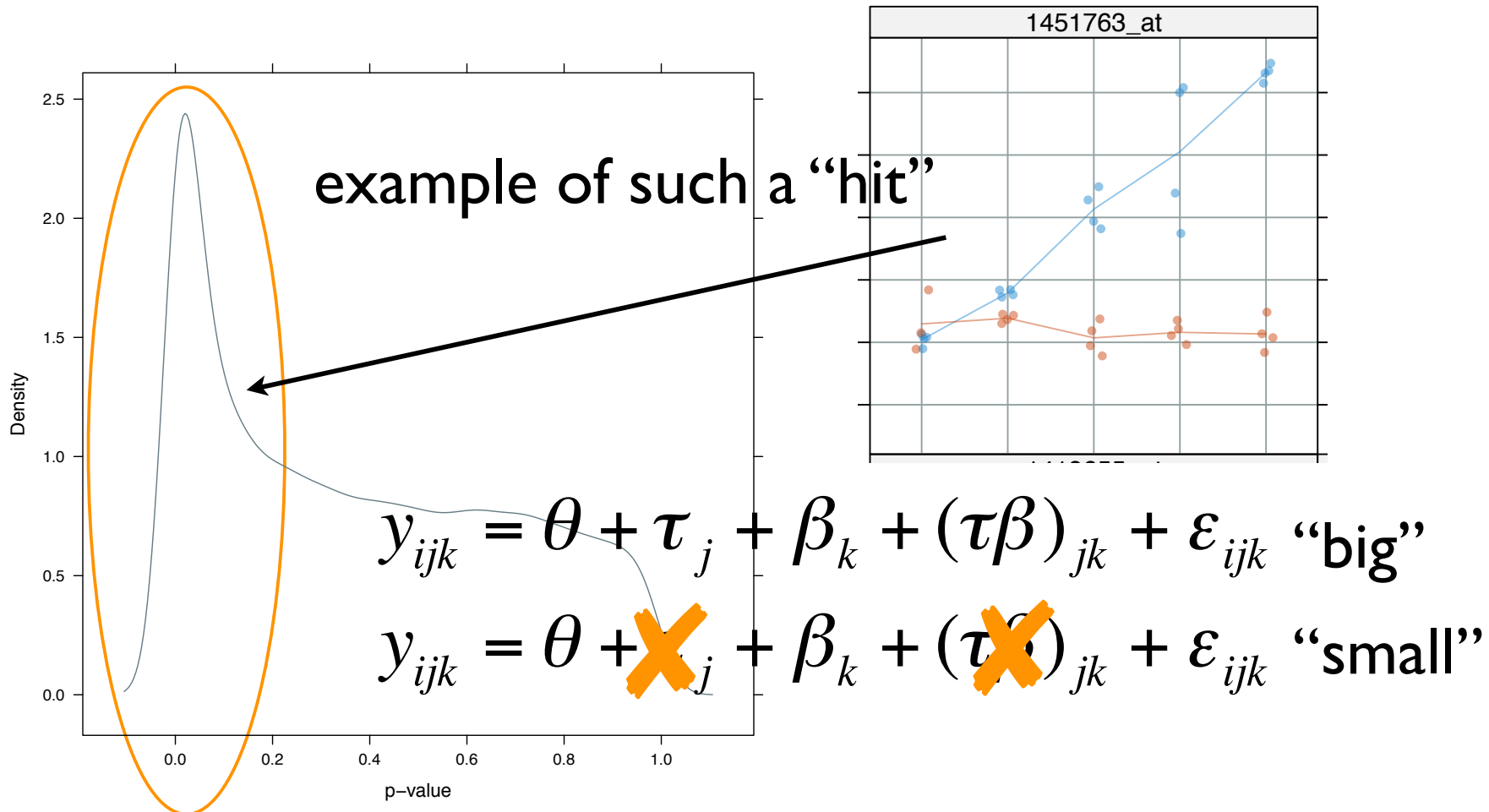
wt      ●  
Nr1KO    ●



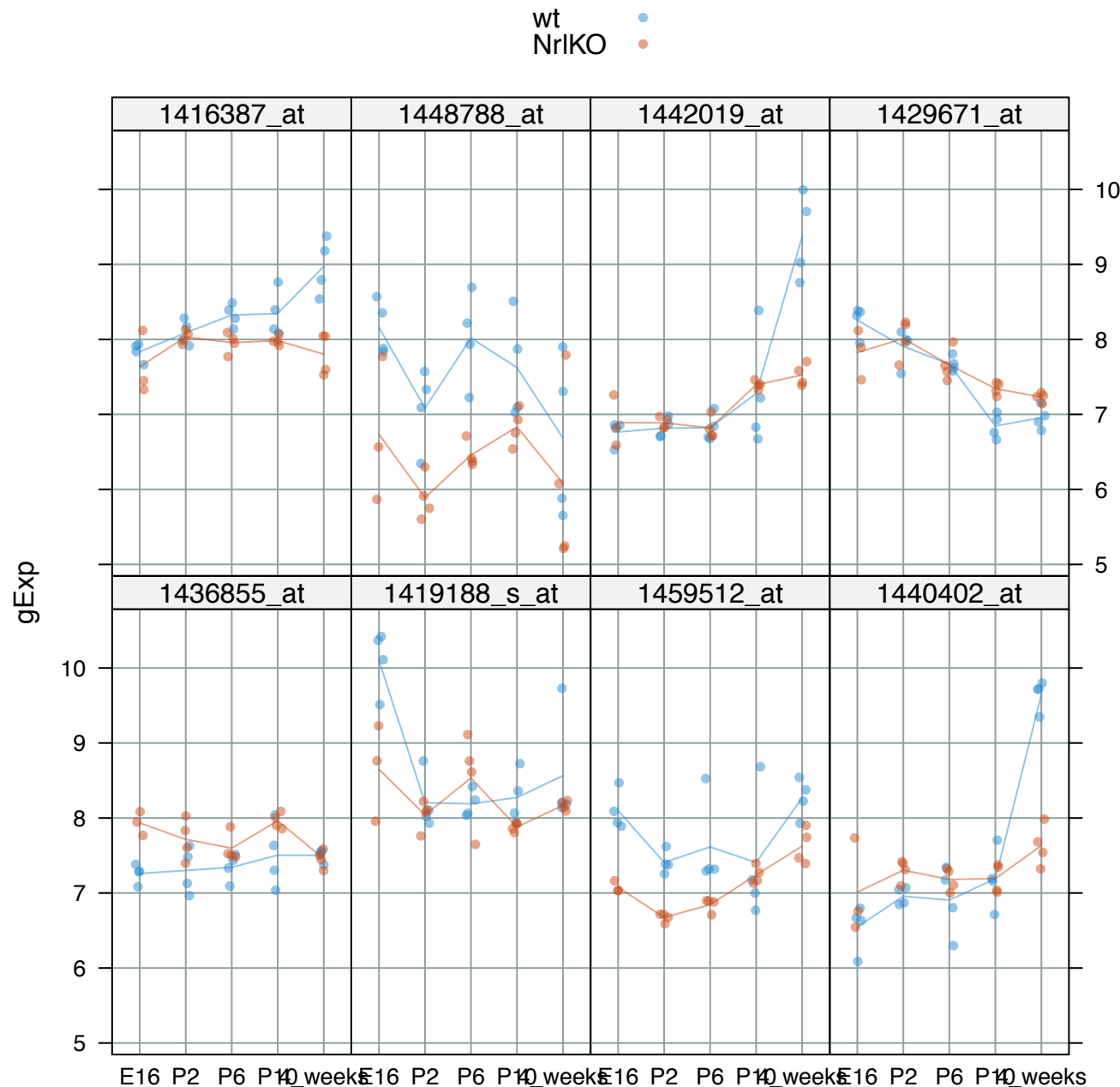
interaction?

# In the two-way ANOVA model, is there evidence that genotype matters? YES.

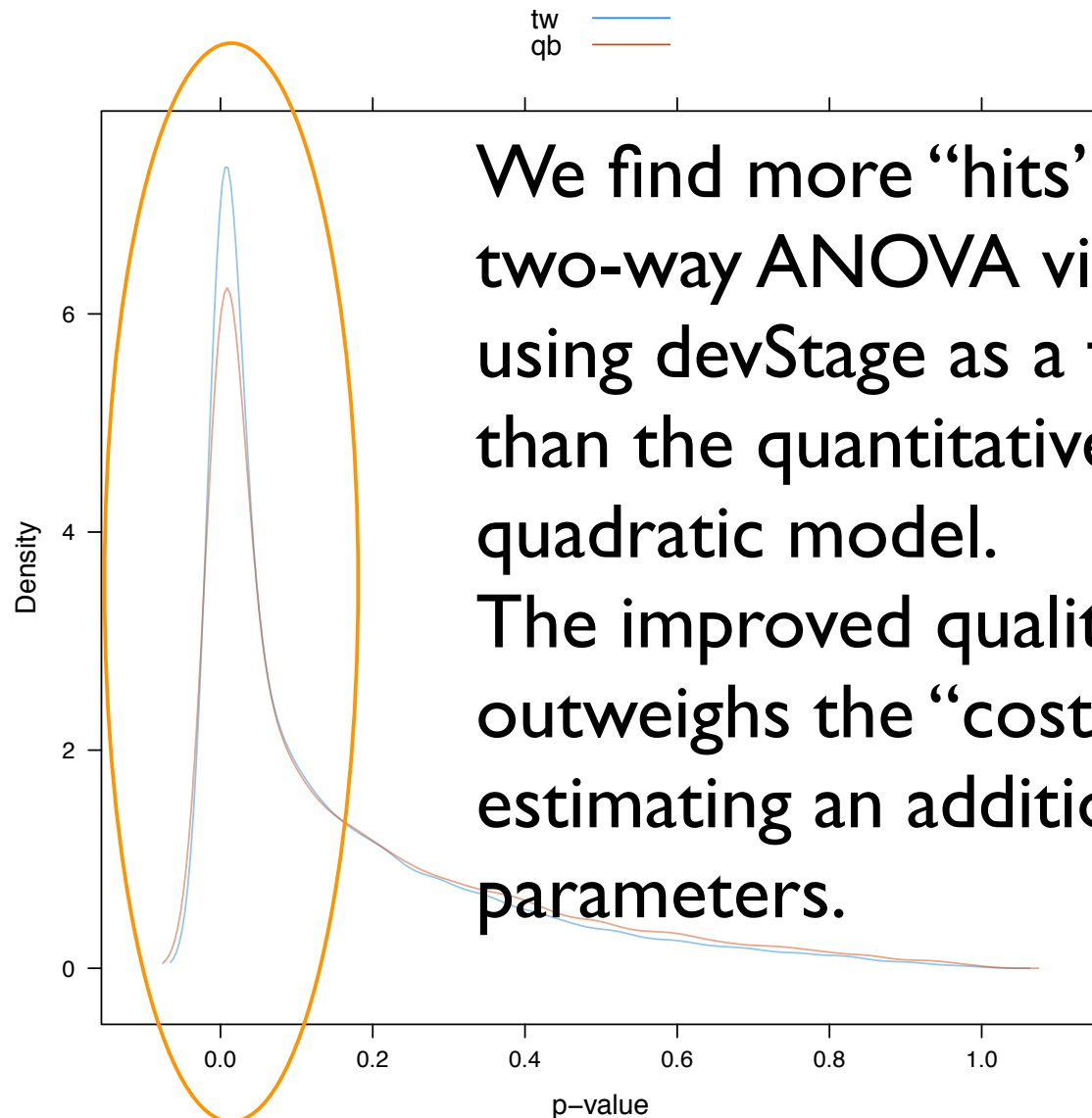
```
## this code is fictional but conveys the point  
anova(lm(gExp ~ gType * devStage), lm(gExp ~ devStage))  
## inspecting the p-values from these F tests
```



# more “gType” hits within the ANOVA models



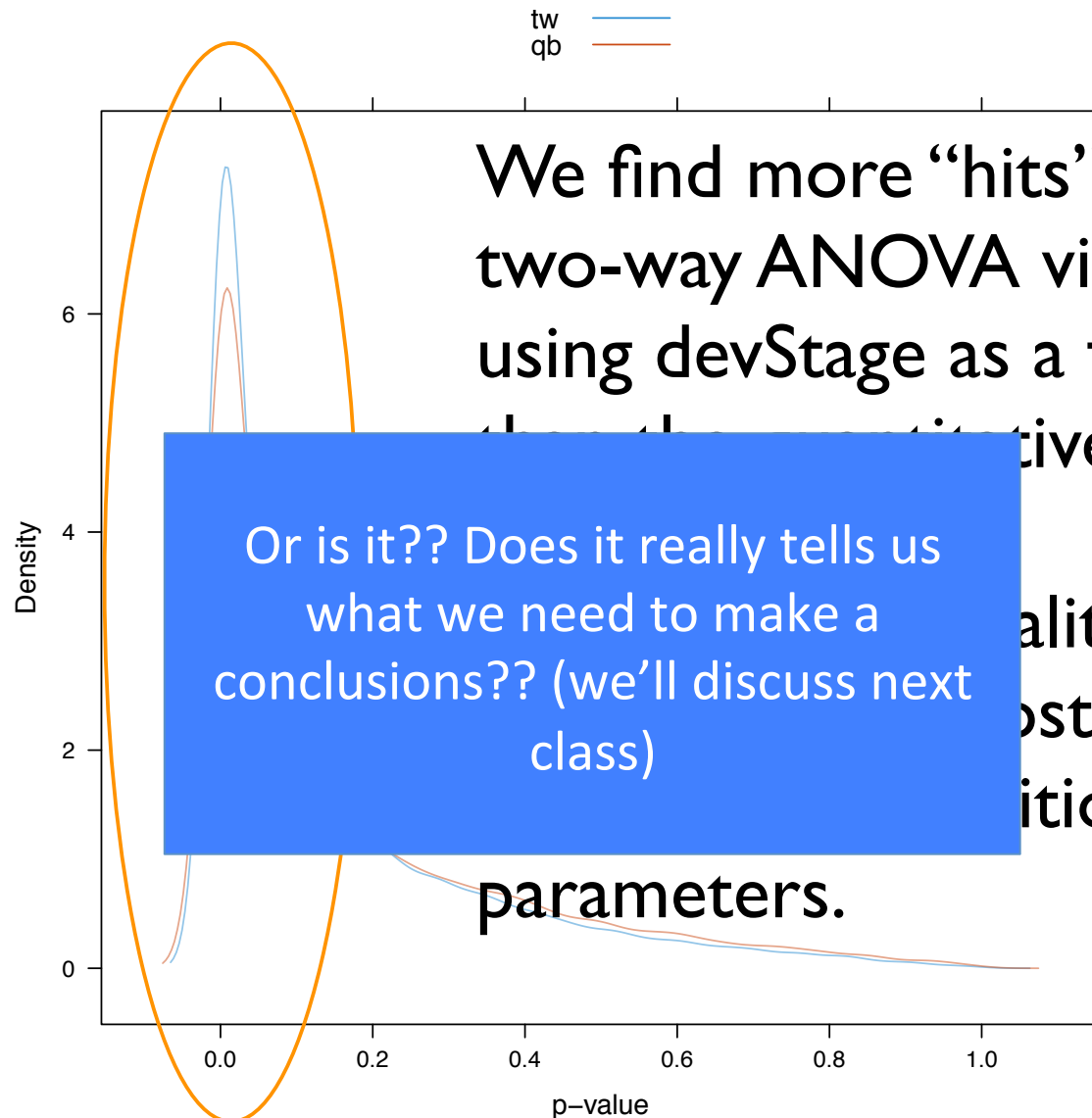
Looking at evidence of any differential expression at all (overall F test) in the two-way ANOVA model vs. the quadratic.



We find more “hits” with the two-way ANOVA viewpoint, i.e. using devStage as a factor rather than the quantitative age and a quadratic model. The improved quality of fit outweighs the “cost” of estimating an additional 4 parameters.



Looking at evidence of any differential expression at all (overall F test) in the two-way ANOVA model vs. the quadratic.



Or is it?? Does it really tells us what we need to make a conclusions?? (we'll discuss next class)

We find more “hits” with the two-way ANOVA viewpoint, i.e. using devStage as a factor rather than devAge and a quadratic term. The quality of fit is “lost” of additional 4 parameters.