Stat 411/511

Assumptions of Regression

Nov 25 2015

Charlotte Wickham

stat511.cwick.co.nz

DA #3 up on website by 3pm

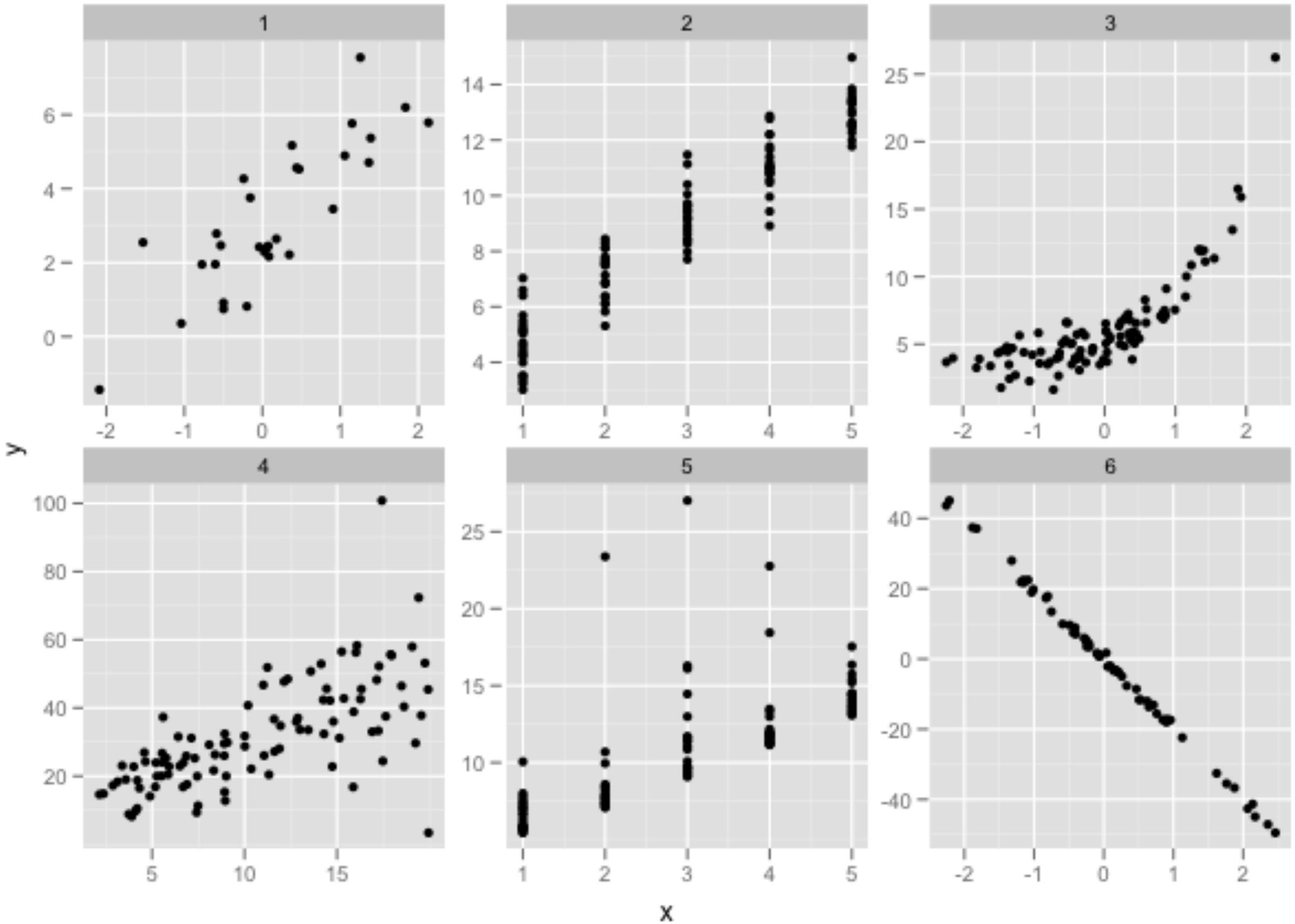
A little more directed

Due next Friday (Dec 4th) @ midnight

Study guide for regression posted

Last year’s final & solution posted

Remember these?



Assumptions

1. **Normal** subpopulation distribution of response at each value of explanatory.
2. The means of the subpopulations fall on a **straight line** function of the explanatory variable. a.k.a. **linearity**
3. The subpopulations have the **same standard deviation**, σ. a.k.a. **constant sd**
4. Observations are **independent**.

The usual analysis procedure

1. Plot response against explanatory

Are there problems that make linear regression look inappropriate? Would a transformation help?

2. Fit linear regression model.

3. Use residuals to reexamine assumptions.

Do the assumptions look good?

Yes, then interpret fit and answer questions of interest (confidence intervals, prediction intervals etc.)

No, then consider transformations, different model for the mean etc.

Three residual plots to check

Residuals against fitted values

Linearity **Constant variance**

Residuals against explanatory variable

Linearity **Constant variance**

Normal probability plot of residuals

Normality

You can also use these to check the assumptions of the one-way ANOVA, if `fit` is your full (or final) model.

Will need the fitted values and residuals

```
fit <- lm(Distance ~ Velocity, data = case0701)
residuals(fit)
fitted(fit)
```

```
# behind the scenes, qplot does
fortify(fit)
```

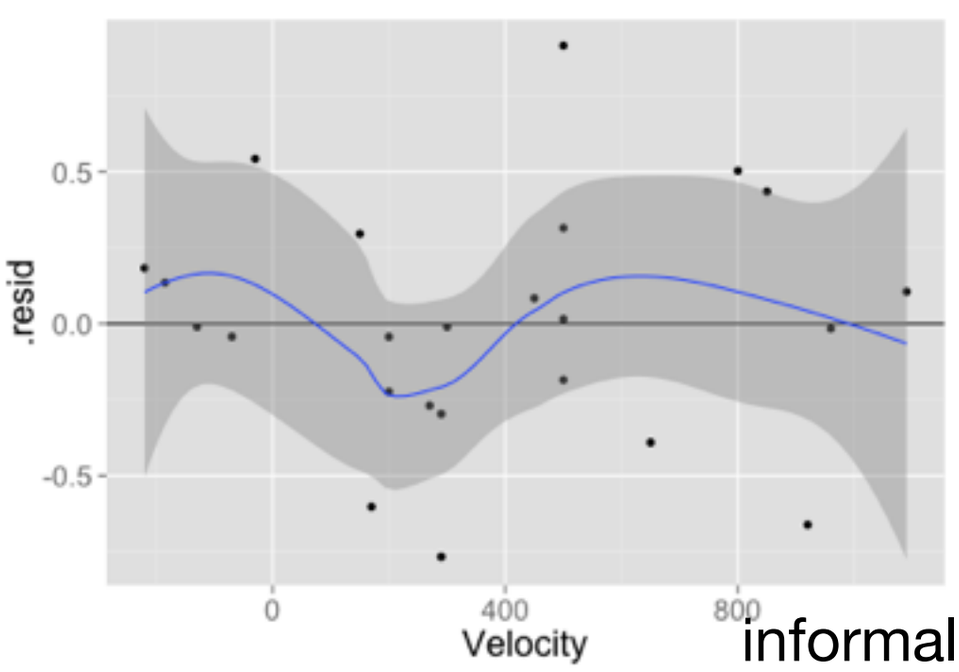
Anova case

```
fit <- lm(Score ~ Handicap, data = case0601)
```

Linearity

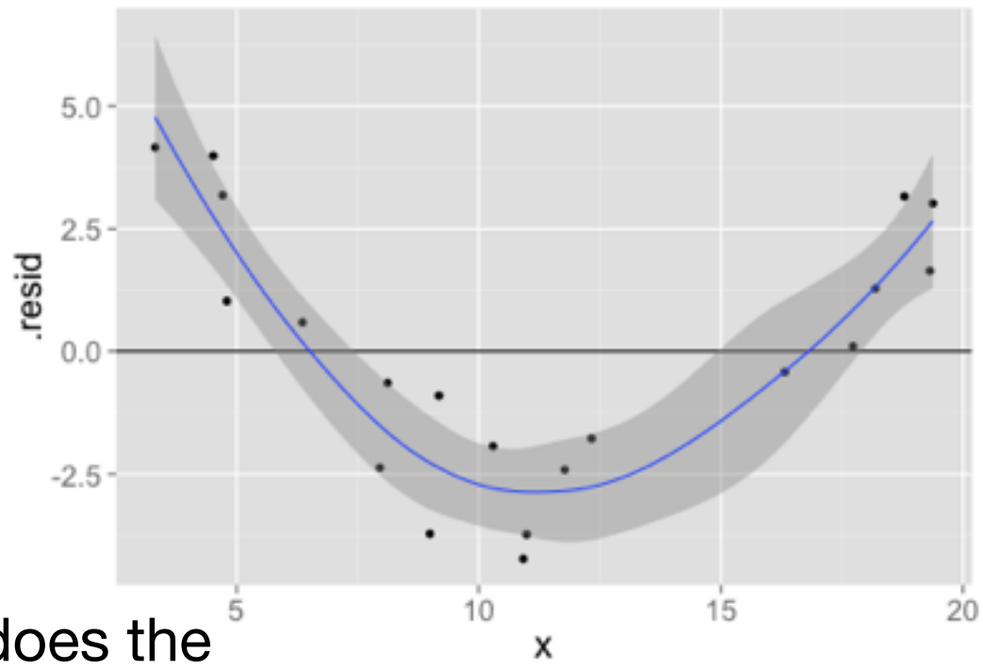
If the linear assumption is true, the residuals should be **equally spread either side of zero**, not systematically above or below zero.

```
qplot(Velocity, .resid, data = fit) +
   geom_hline(yintercept = 0) +
   geom_smooth()
```



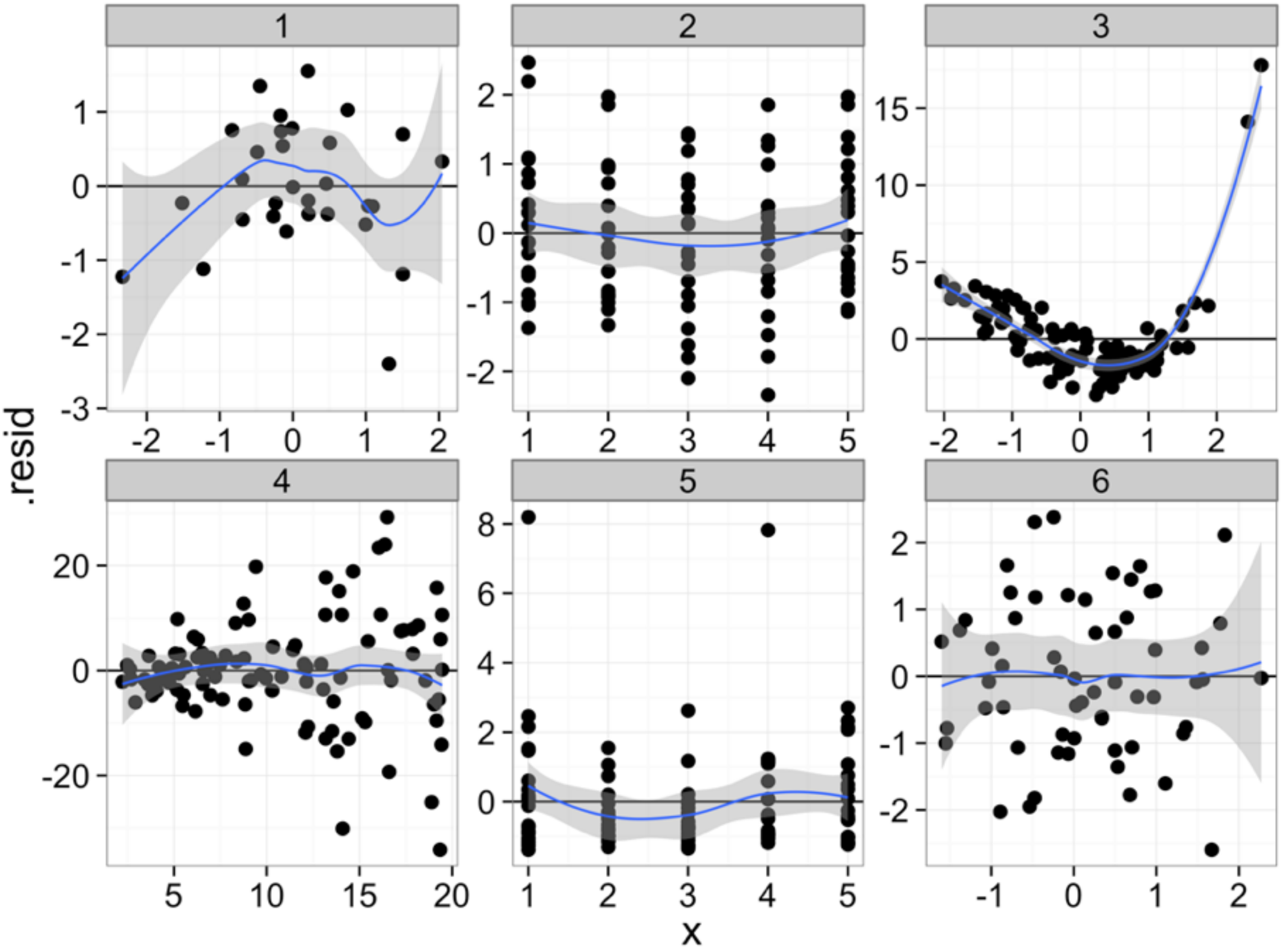
OK!

informally, does the flat black line lie within the grey band



Not OK!

Are there systematic deviations from a flat line?



Linearity

Robustness: Not robust. The least squares estimates **will be biased**.

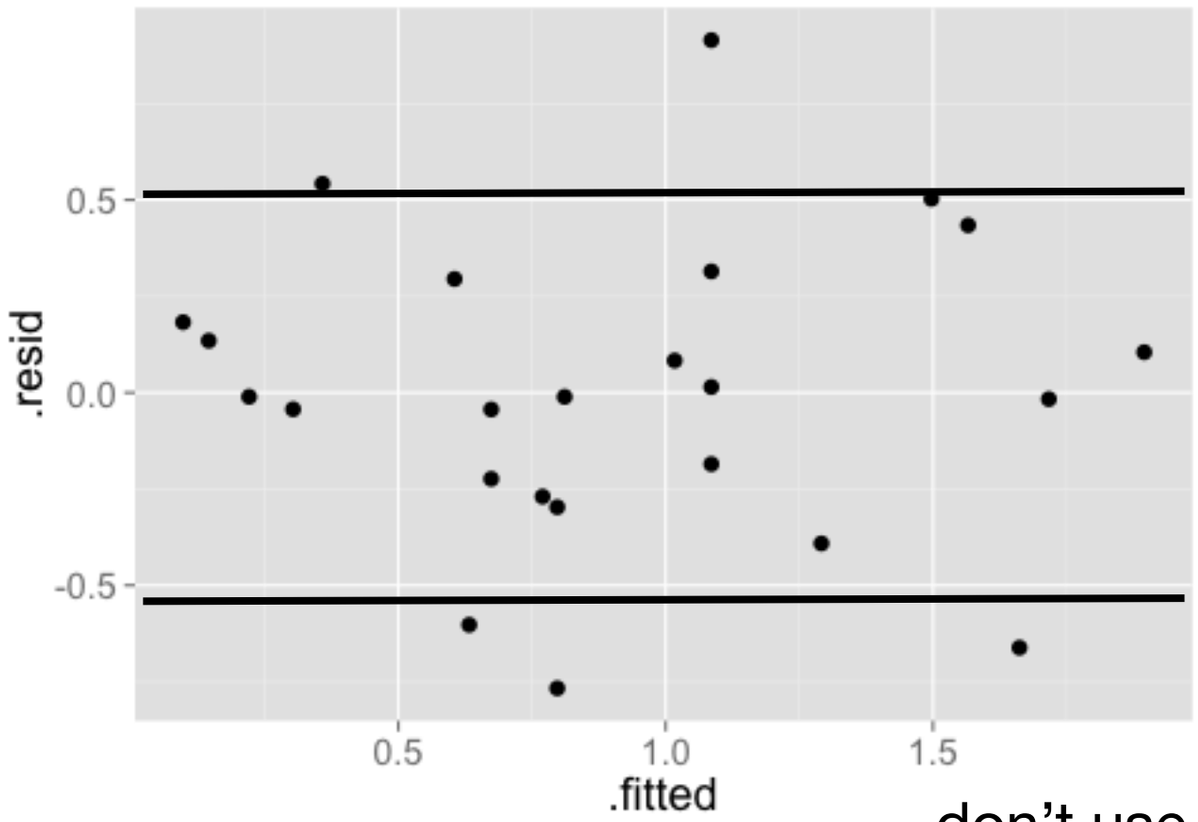
Inference will be misleading.

Remedy: Consider a more complicated model for the mean, or transform response, or explanatory or both.

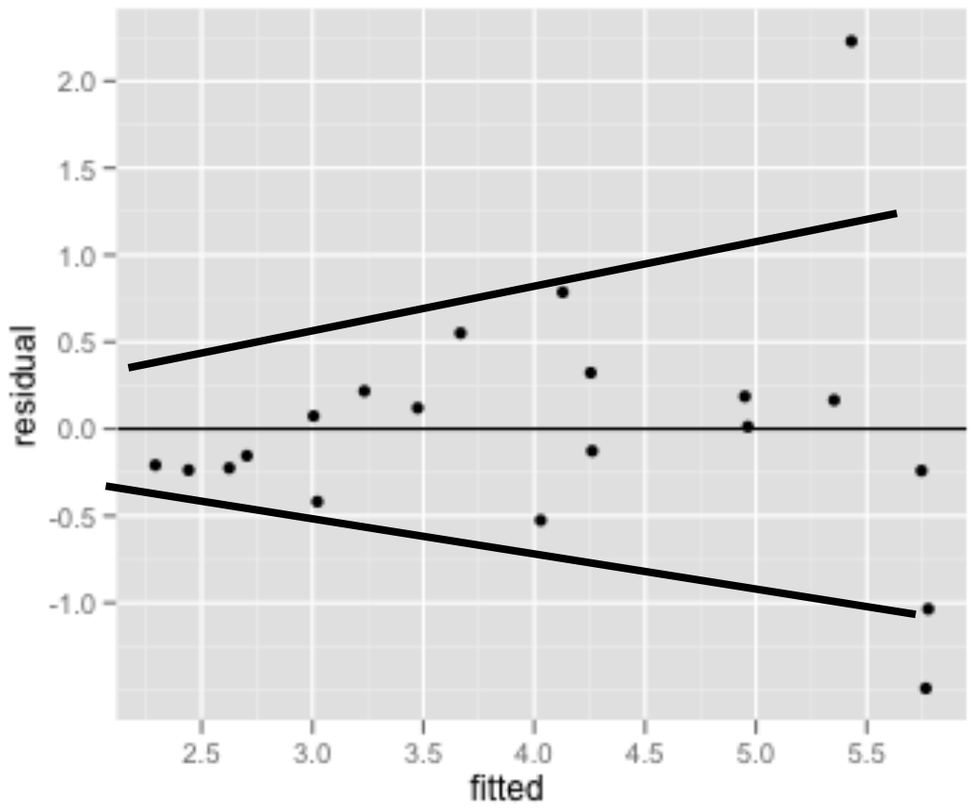
Constant spread

If the equal SDs assumption is true, the residuals should show **constant spread** at all values of the explanatory.

```
qplot(.fitted, .resid, data = fit)
```



OK!



Not OK!

don’t use the grey band to evaluate this!

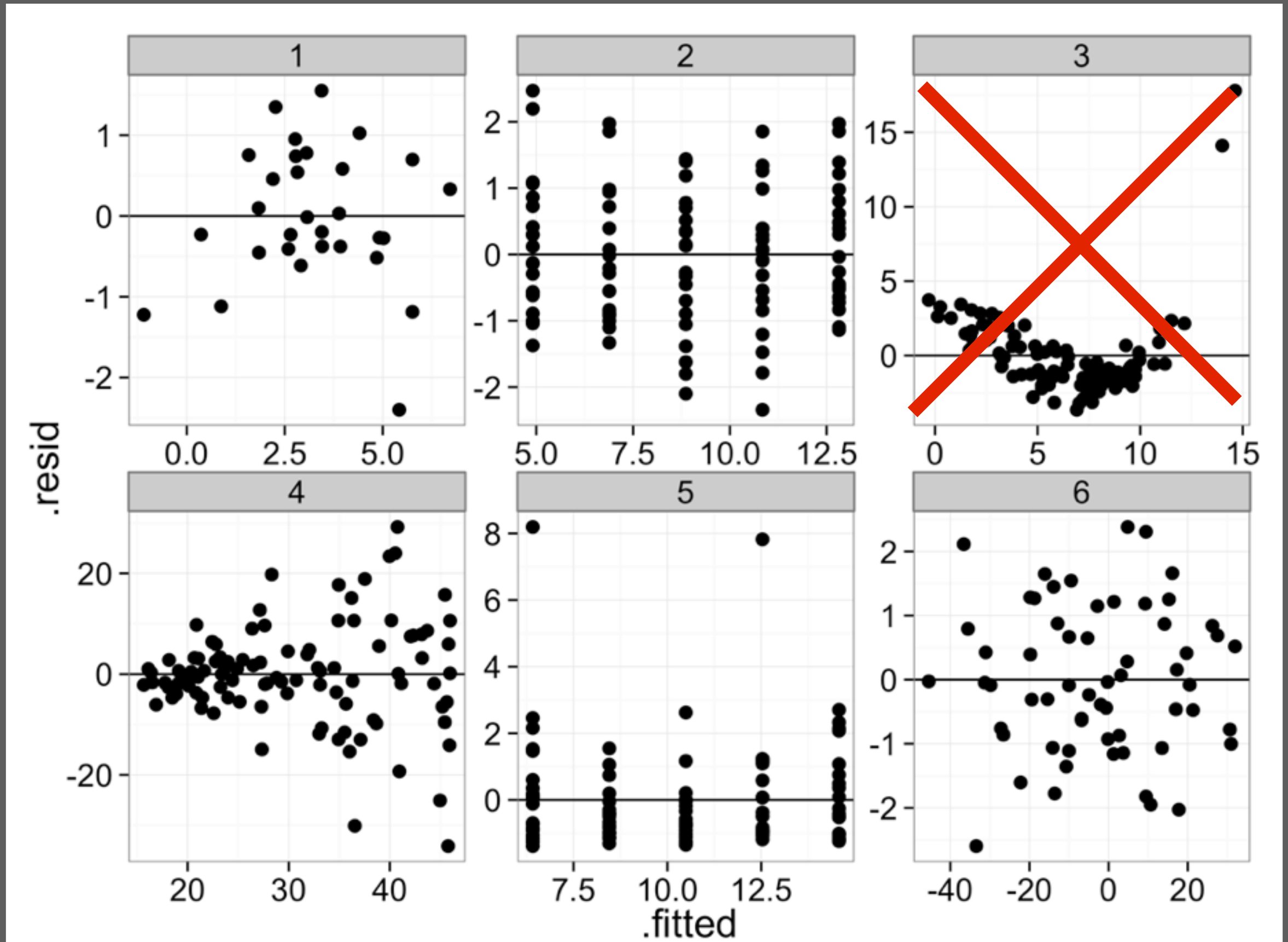
Constant spread

Robustness: Not robust. The least squares estimates will still be unbiased, but the standard errors will be wrong.

I.e. $\hat{\beta}_0$ and $\hat{\beta}_1$ will be good estimates, but your confidence intervals and tests will be inaccurate.

Remedy: Consider transforming the response.

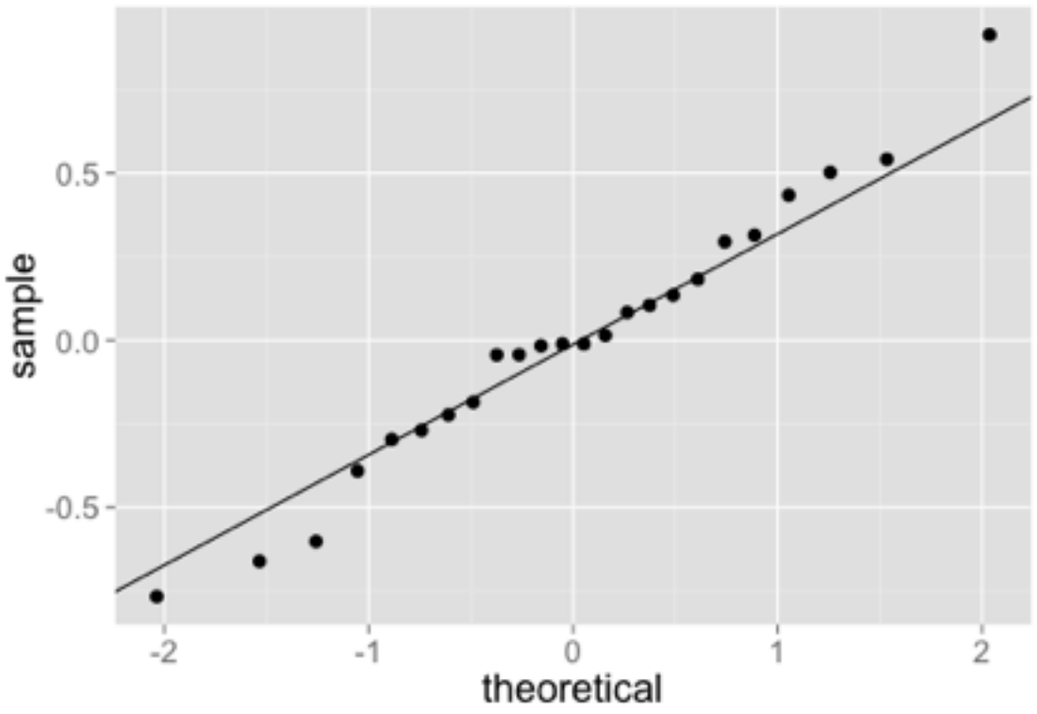
Does the spread look constant?



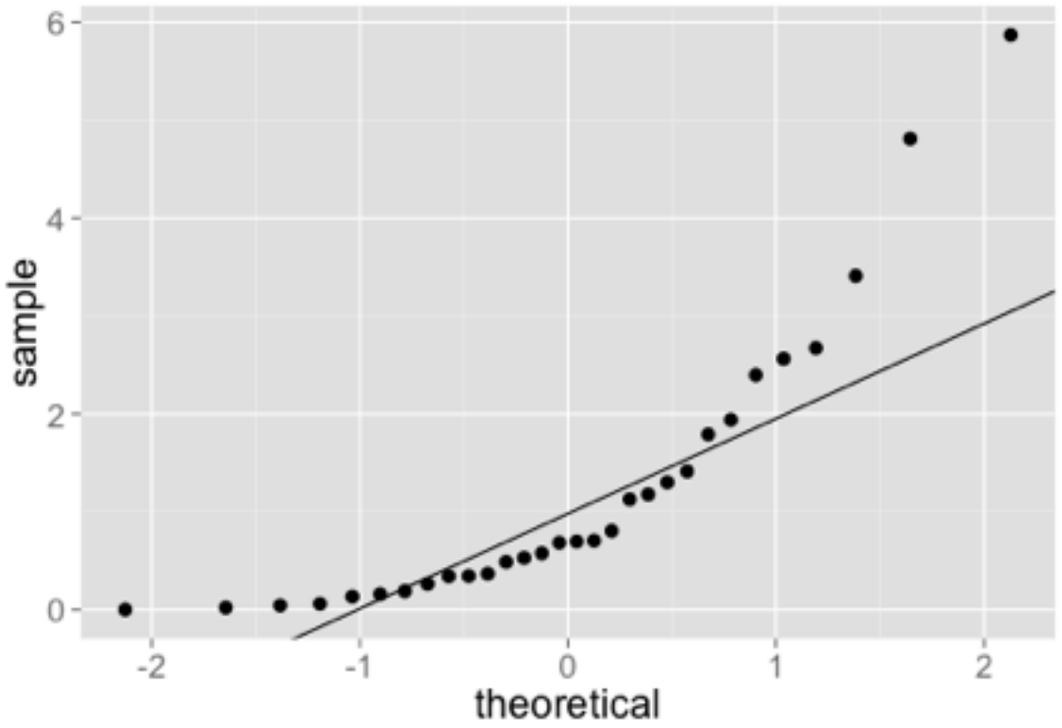
Normality

If the subpopulations are normally distributed the **residuals** should be approximately normally distributed. A normal probability plot of the residuals should show a **straight line**.

```
qplot(sample = .resid, data = fit)
```



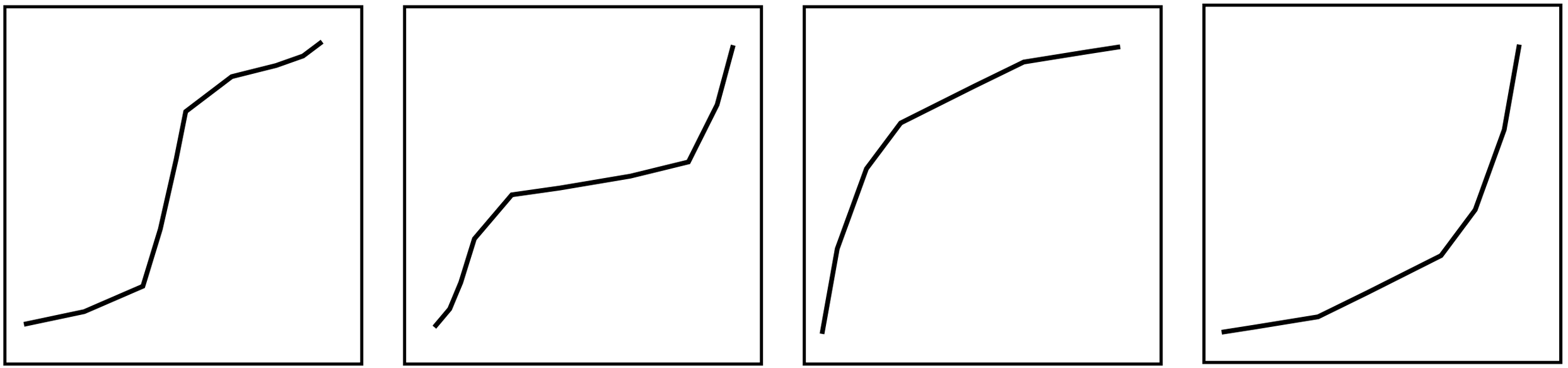
OK!



Not OK!

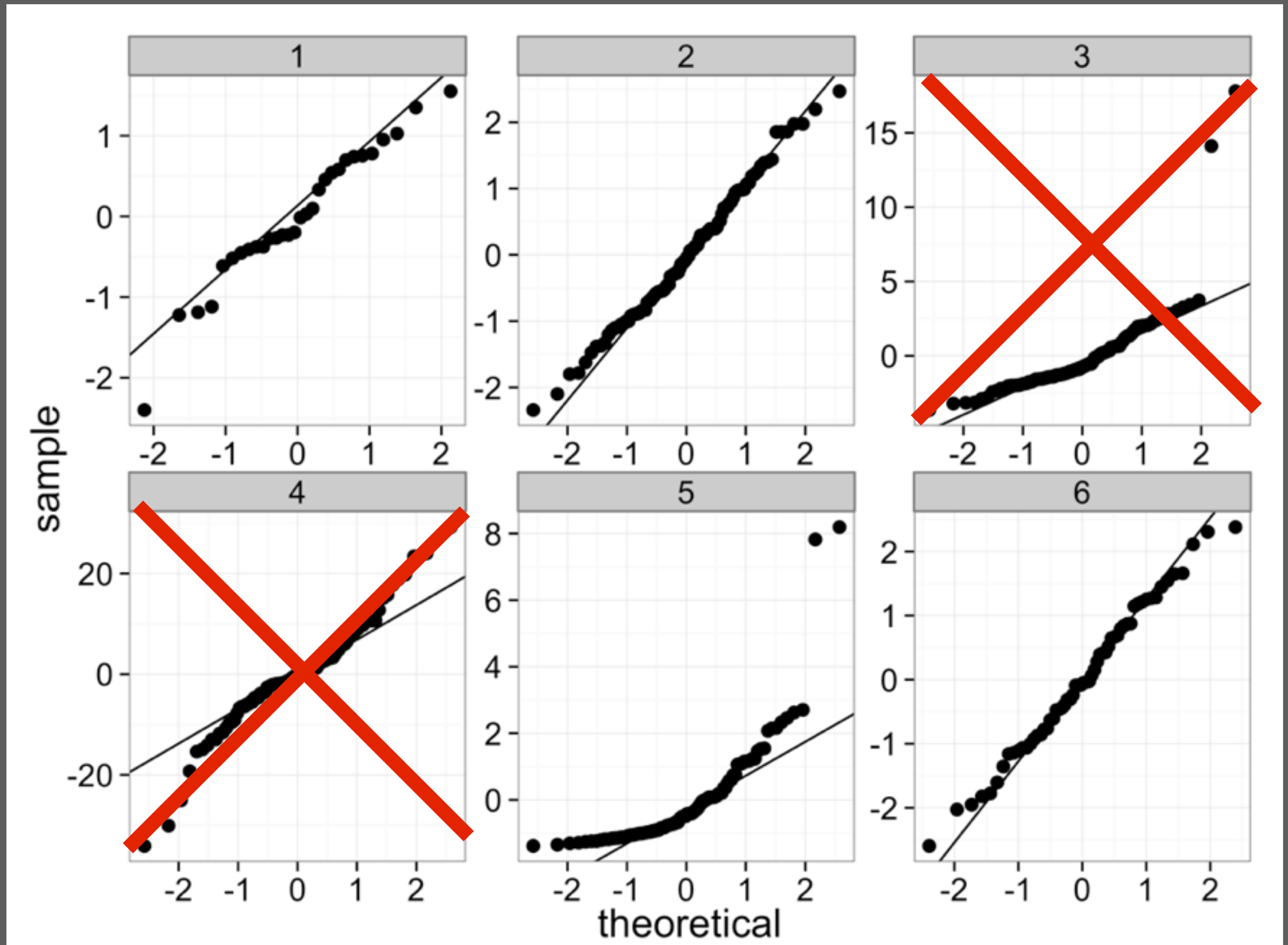
Normal probability plot

Common bad shapes

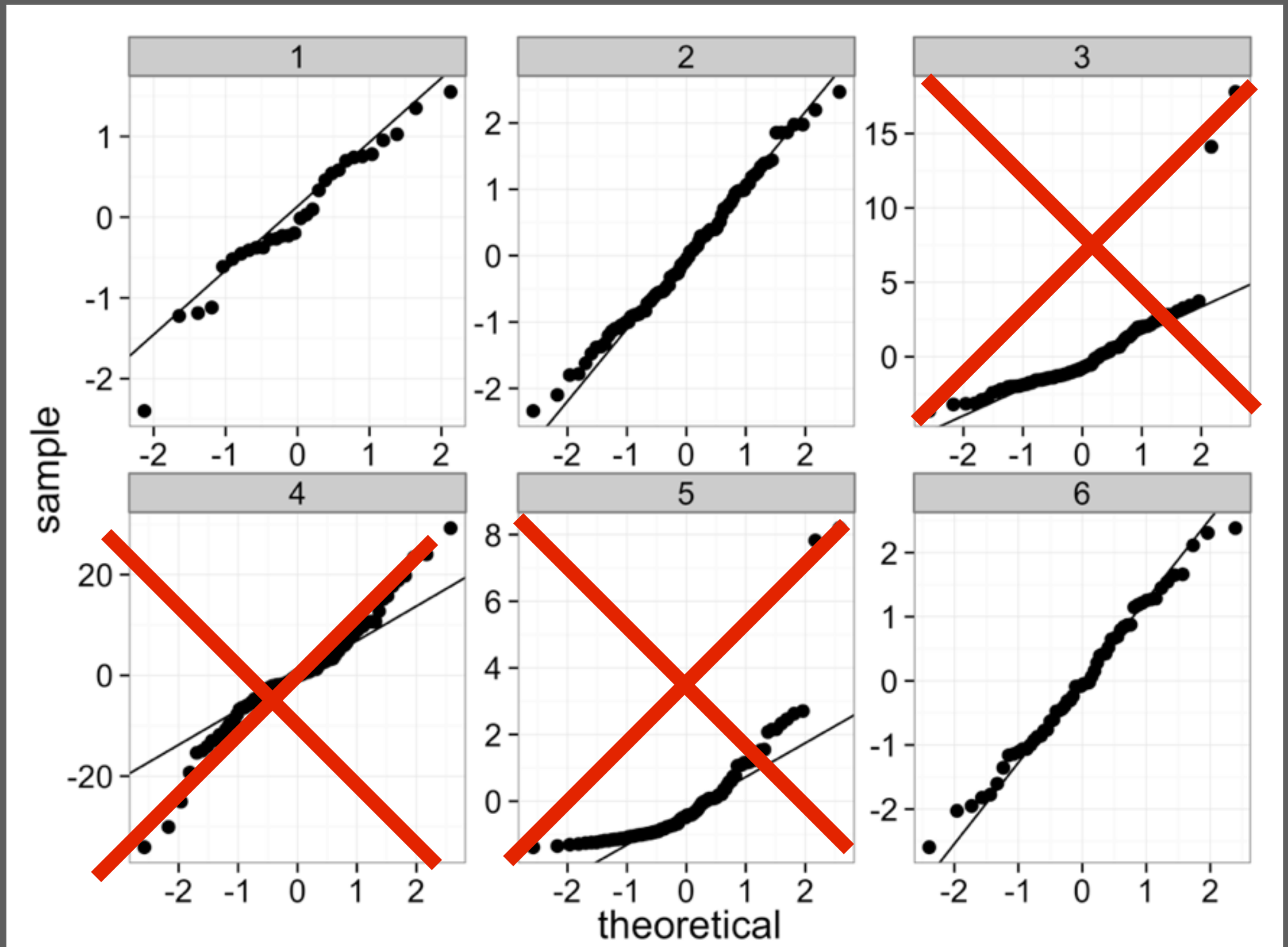


All not OK!

Do the points fall on a straight line?



Do the points fall on a straight line?



Normality

Robustness: Robust with large samples. The sampling distributions of least squares estimates will be approximately normal (CLT again).

Prediction intervals can be misleading.

Remedy: Transformation of response.

Or generalized linear models (ST513)

Independence

Robustness: Not robust. The least squares estimates will still be unbiased, but the standard errors will be wrong.

Remedy: More complicated models.

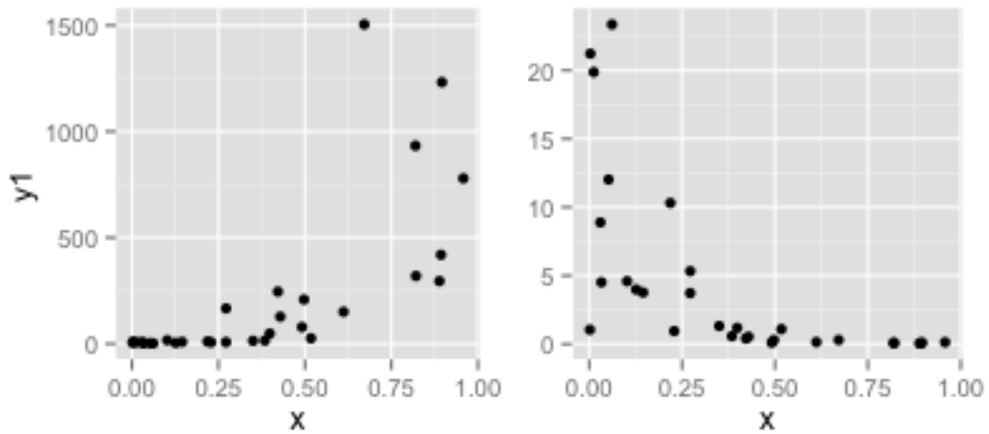
(ST512)

Log transforms in simple linear regression

Three possibilities

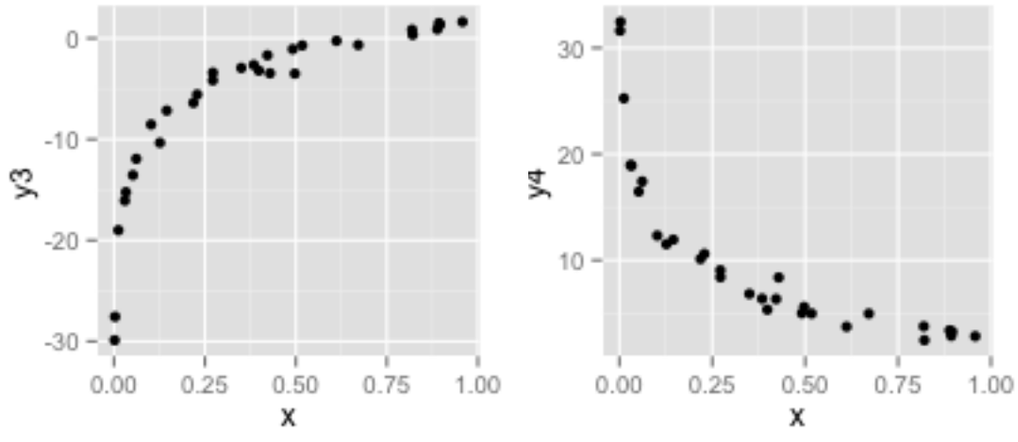
Log transform Y

spread is larger with larger μ{Y},
and μ{Y|X} doesn't look linear in X.



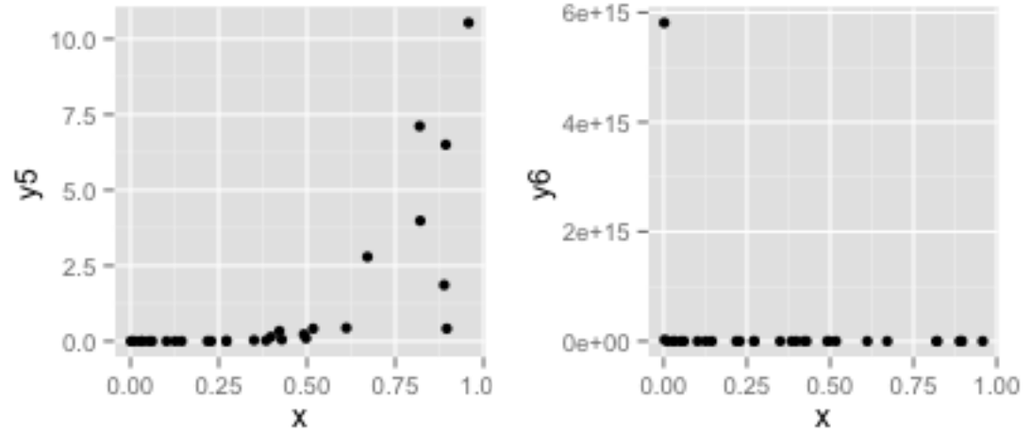
Log transform X

spread is constant, but
μ{Y|X} doesn't look linear in X.



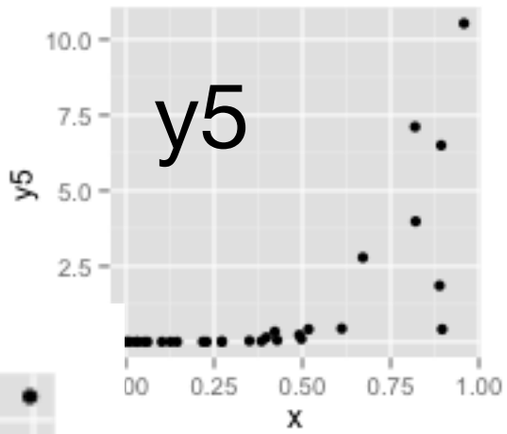
Log transform X and Y

spread is larger with larger μ{Y},
and μ{Y|X} doesn't look linear in X.

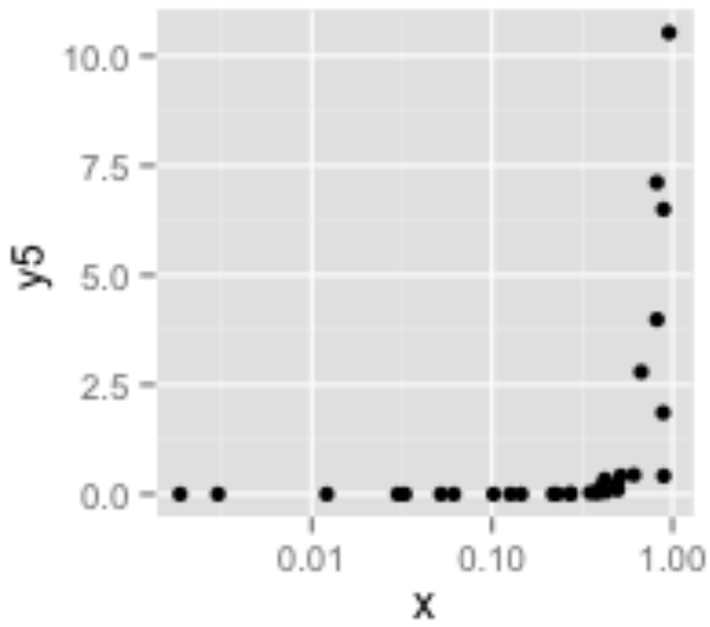


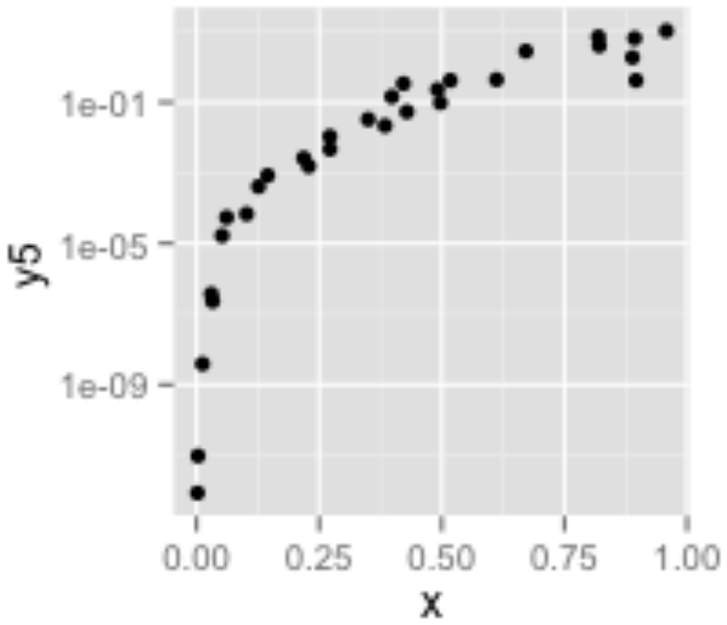
Easiest way to tell...

Try plotting on a log scale to achieve linearity and constant variance.



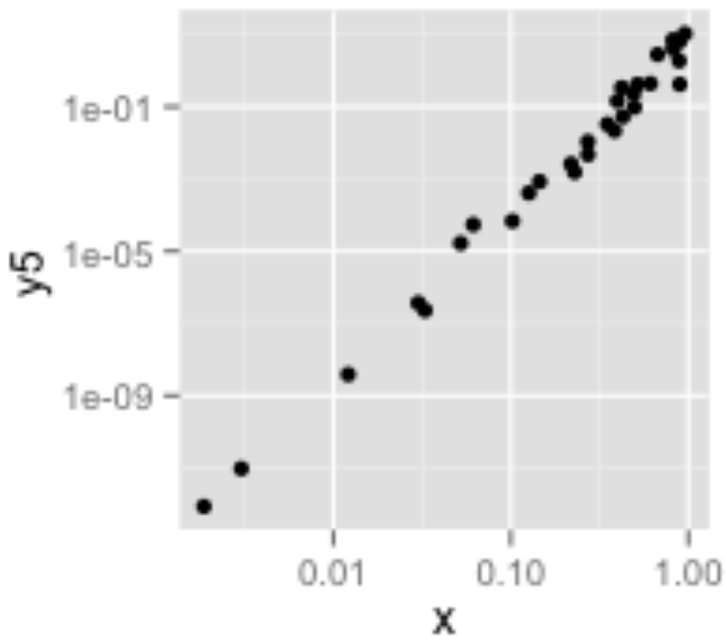

```
qplot(x, y5, data = df_log, log = "x")
```






```
qplot(x, y5, data = df_log, log = "y")
```

```
qplot(x, y5, data = df_log, log = "xy")
```



Not examinable

Interpretation of slope in the three possibilities

Log transform Y

A **1-unit** increase in X is associated with a **multiplicative** increase of **exp(β_1)** in the **median**.

assuming mean Y equals median Y on transformed scale

Log transform X

A **doubling** of X is associated with an **additive** increase of **$\beta_1$log 2** units in the **mean** response.

Log transform X and Y

A **doubling** of X is associated with a **multiplicative** increase of **2^{β_1}** times in the **median** response.

assuming mean Y equals median Y on transformed scale

Happy Thanksgiving!