# Principles of model selection

# Linear models as tests of treatment effects

- We often think of statistical analysis as tests of a treatment effect
  - If the treatment has no effect at all the null hypothesis would be true (Ho: $\mu_{Treatment} = \mu_{Control}$)
  - Testing the null hypothesis is a test of the treatment effect
- Good way to think about statistical analysis for simple experiments:
  - Variable Treatment, levels Treatment and Control
  - Single measured response
  - All nuisance variables held constant
  - Random assignment of subjects to treatment levels

# Best case: complete, balanced designs

- Experiment to find the conditions under which potatoes rot the slowest
- Response variable is a measure of rot
- Factors tested that could affect rotting speed were:
  - BAC = bacterial inoculation (3 levels)
  - TEMP = temperture (2 levels)
  - OXYGEN = oxygen levels (3 levels)
- Question is: which combination of factors give the lowest amount of rot?

```
Response: ROT
                Df Sum Sq Mean Sq F value    Pr(>F)
BAC              2 651.81  325.91 13.9123 3.339e-05 ***
TEMP             1 848.07  848.07 36.2024 6.599e-07 ***
OXYGEN           2  97.81   48.91  2.0877   0.13872
BAC:TEMP         2 152.93   76.46  3.2640   0.04981 *
BAC:OXYGEN       4  30.07    7.52  0.3209   0.86207
TEMP:OXYGEN      2   1.59    0.80  0.0340   0.96661
BAC:TEMP:OXYGEN  4  81.41   20.35  0.8688   0.49206
Residuals       36 843.33   23.43
```

# It's easy when design is orthogonal

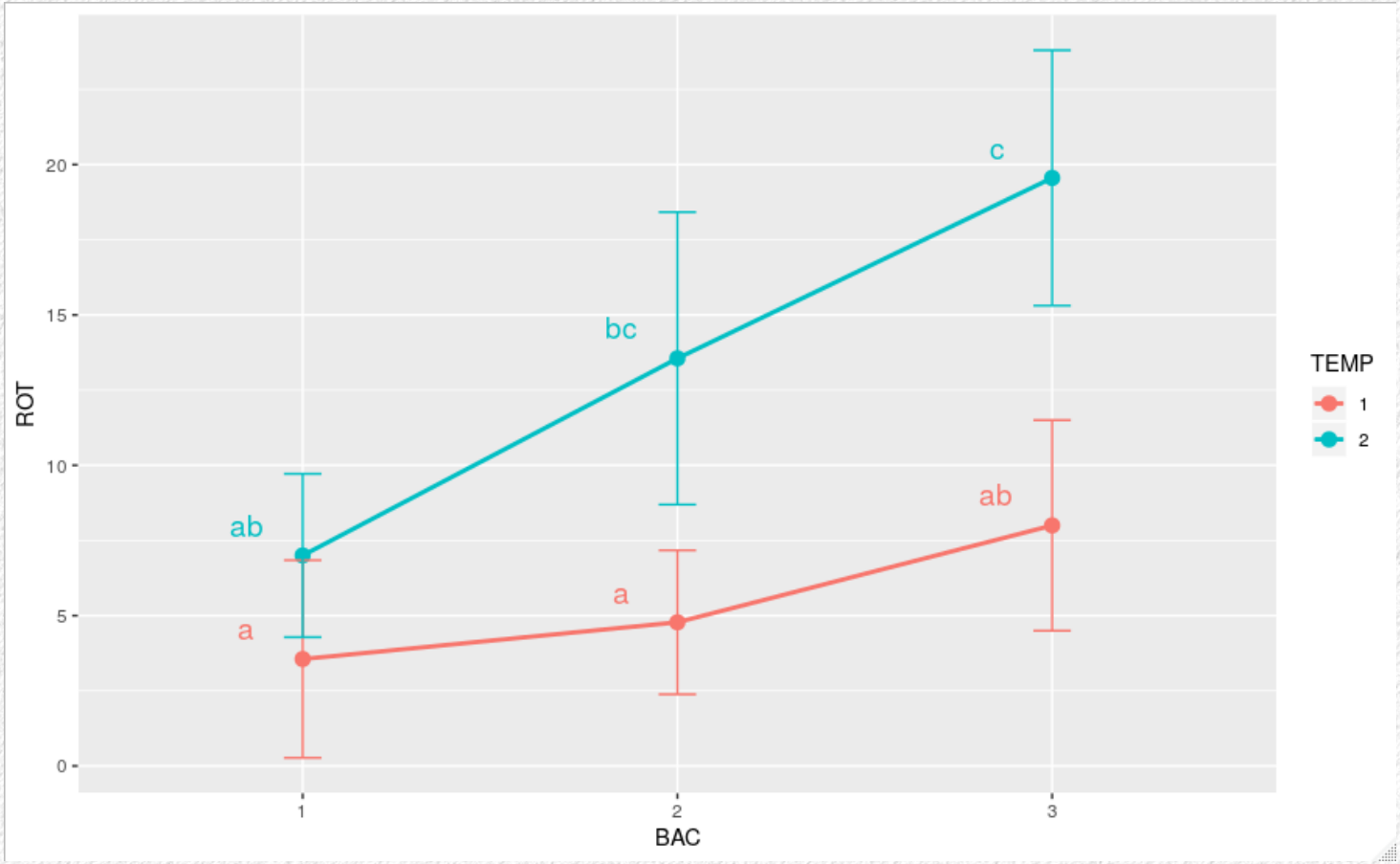
*Include all the predictors initially*

```
Response: ROT
           Df  Sum Sq Mean Sq F value    Pr(>F)
BAC         2  651.81  325.91 14.8390 9.608e-06
TEMP        1  848.07  848.07 38.6138 1.180e-07
BAC:TEMP    2  152.93   76.46  3.4815   0.03874
Residuals  48 1054.22   21.96
```

*Drop all non-significant terms*

*SS are the same, but p-values are smaller, why?*

# Interpreting the best model

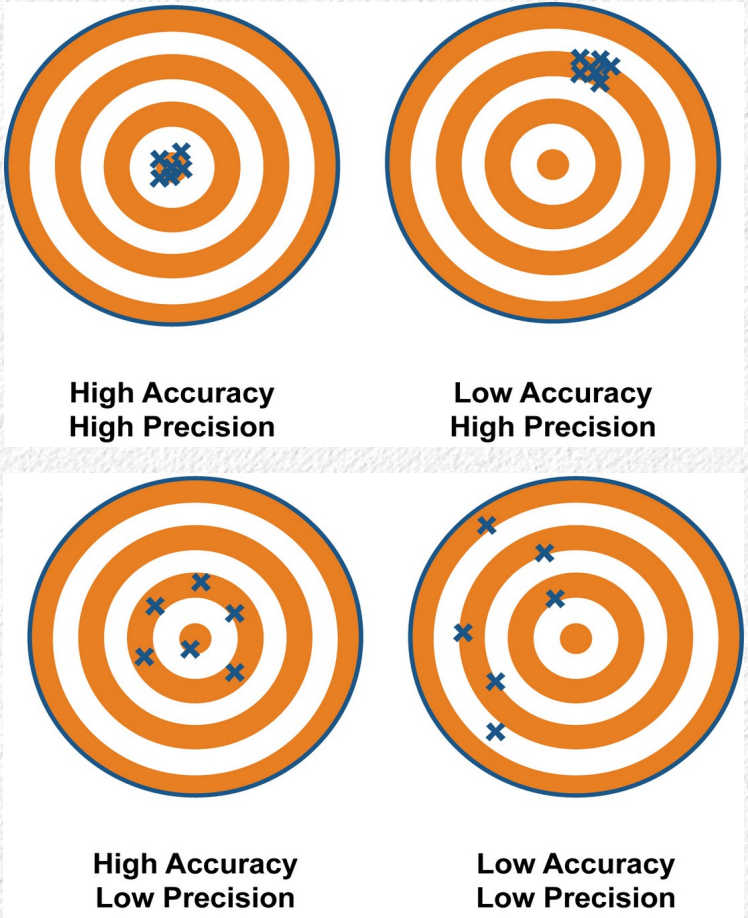


# But, studies are often not so simple

- Complex experimental conditions
  - Nuisance variables (some of which can't be randomly assigned)
  - Covariates (some of which can't be blocked) → lack of independence of predictors
  - Repeated measurements of the same individuals → lack of independence of data values
- Complex responses to treatment variables
  - Possible interactions, nonlinearities
  - Multiple correlated predictors, confounding
- Best to approach analysis of complex studies as statistical modeling of the structure in the data
  - Once the best model is found, it can be interpreted

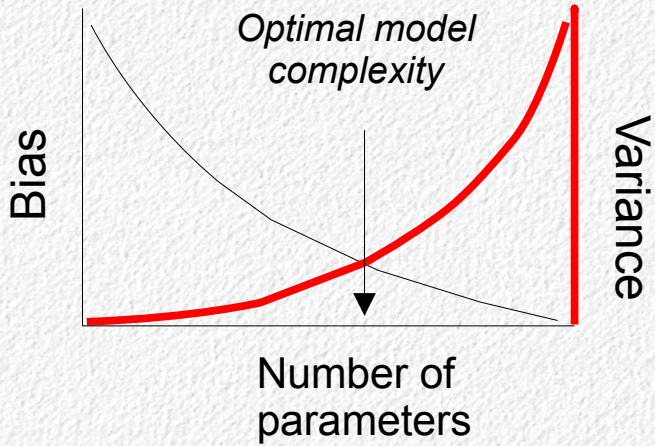
# Model bias and variance

- Bias = an estimator does not equal the parameter it estimates over the long run
  - Opposite of accuracy
  - Bias assessed by whether the model puts predicted values in the middle of the data
- Variance = how far apart repeated estimates are from one another
  - Opposite of precision
  - Standard error of estimates represents variance



# Problem: bias and variance trade off

- We can decrease bias by making a model more complex
  - Adding variables
  - Adding interactions
  - Adding quadratic, cubic, etc. terms
- Doing this increases variance because:
  - Predictors reduce error DF
  - If predictors are correlated standard errors increase
- Building statistical models balances bias and variance



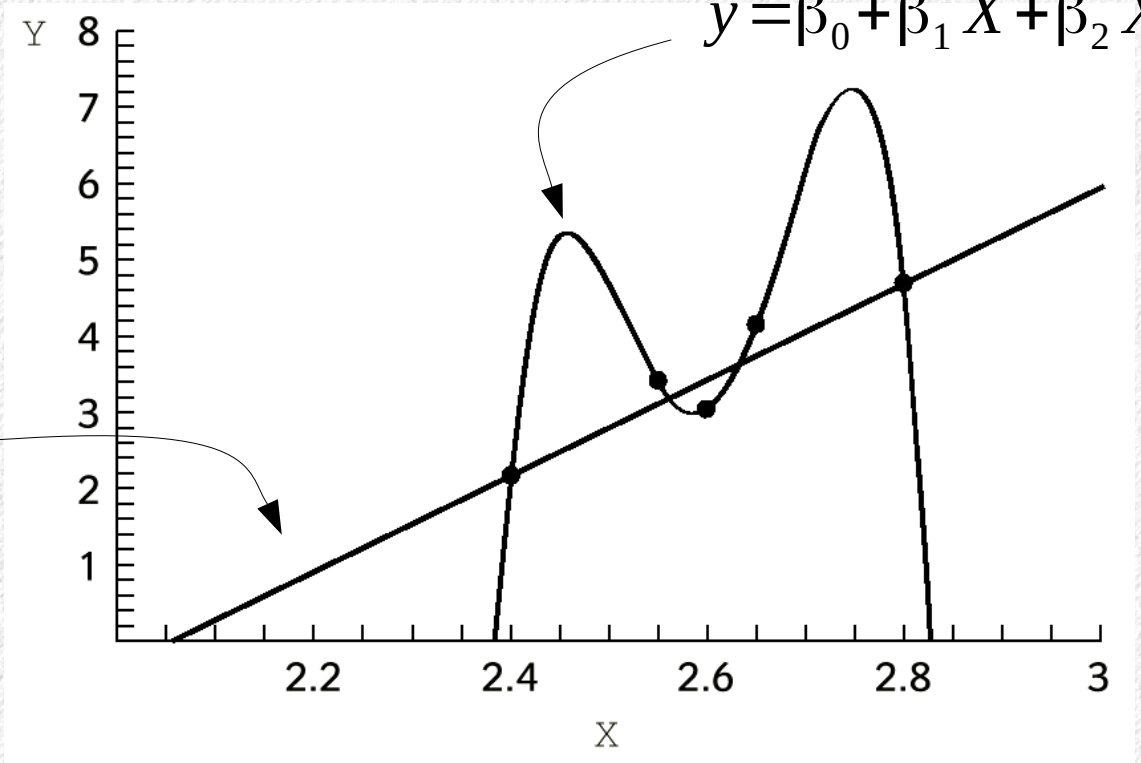
# Problem: complex models don't generalize well

- Seems that we should want to explain as much variation in the response as possible
  - 100% explained variation is the goal
  - More explained variation seems to indicate better understanding of the underlying causes of variation in the response
- But, need to prevent over fitting
  - Over-fitted model is tailored to the quirks of the data set on which it was developed → high $R^2$
  - Incorrectly attributing some of the random variation to real, fixed effects
- An over-fitted model performs poorly when it's applied to a new data set → doesn't generalize well

# Example – which line is a better model for the data?

$\hat{y} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4$

$\hat{y} = \beta_0 + \beta_1 X$

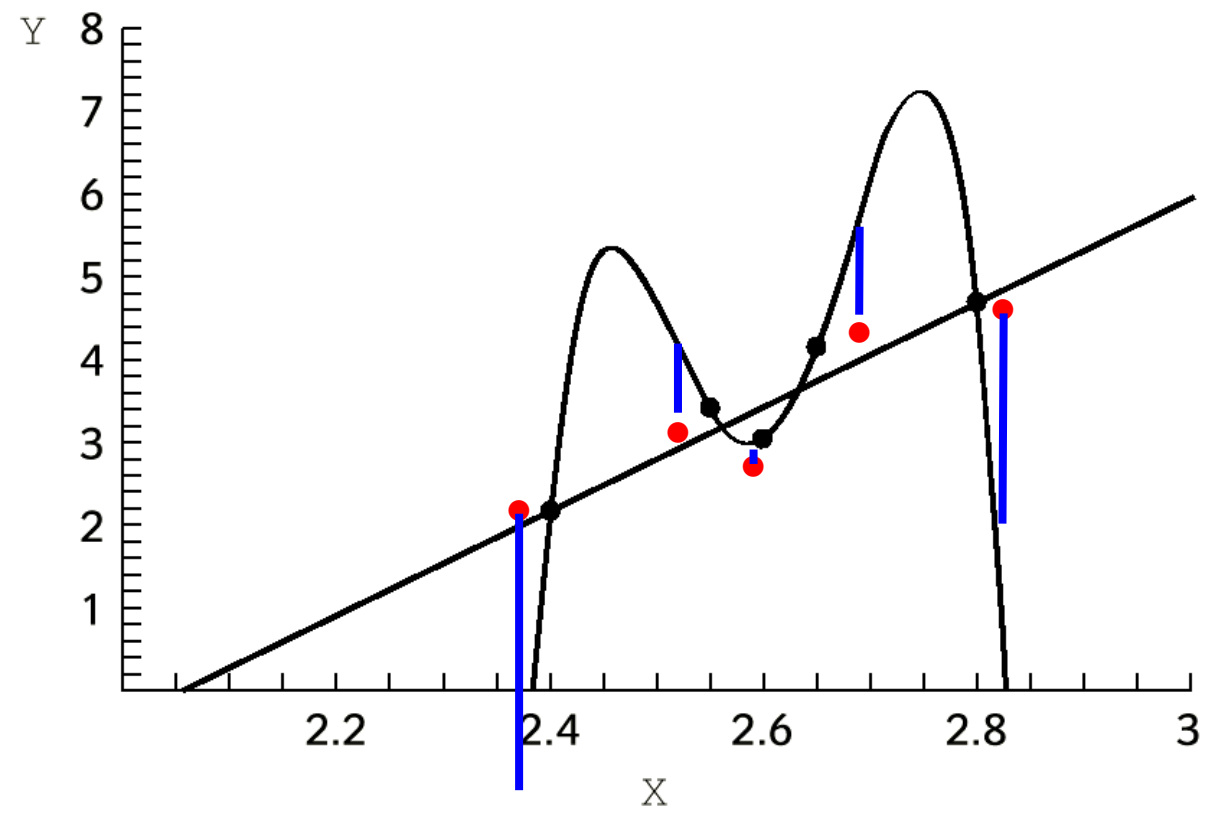



*Which line has the highest $r^2$?*

# Which looks better now?

*Small changes in position of data points ruin the polynomial $R^2$*

*Nearly the same for the straight line*

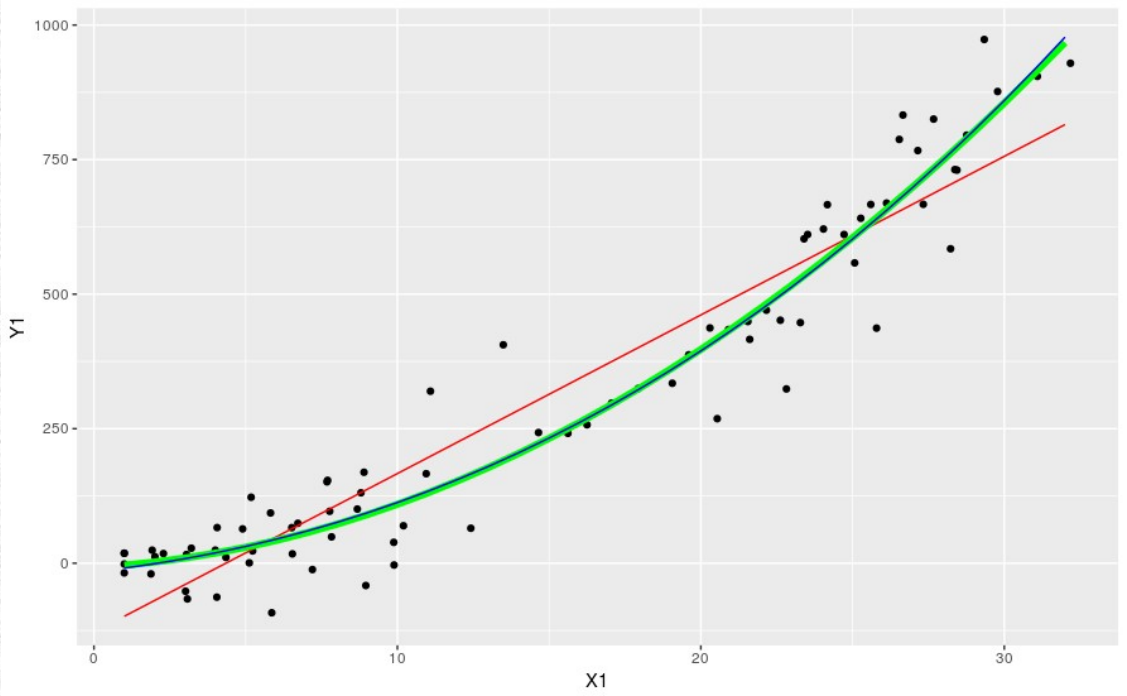



*Even though the polynomial fits the original data better, the simple linear model generalizes better*

# What is the right amount of complexity?

- The number of models possible is often large, increases rapidly with number of predictors
  - With 3 predictors there are 7 models without interactions, 15 with interactions
  - With 6 predictors there are 63 models without interactions, more than 30,000 with interactions
- Q: What is the minimum acceptable model complexity?
- A: Whatever is needed to meet model assumptions

# Minimally, must meet assumptions




First rule is that the model must meet GLM assumptions – as complex as needed to do this

*For continuous predictors, consider adding polynomial terms, if there is evidence of non-linearity*

*Can we meet GLM assumptions with a linear, quadratic, or cubic fit?*

General Linear Model

Word equation: Y1 = X1

X1 is continuous

Analysis of variance table for Y, using Adjusted SS for tests

| Source | DF | Seq SS | Adj SS | Adj MS | F | P |
|---|---|---|---|---|---|---|
| X1 | 1 | 6 663 021 | 6 663 021 | 6 663 021 | 722.89 | 0.000 |
| Error | 78 | 718 946 | 718 946 | 9 217 | | |
| Total | 79 | 7 381 967 | | | | |

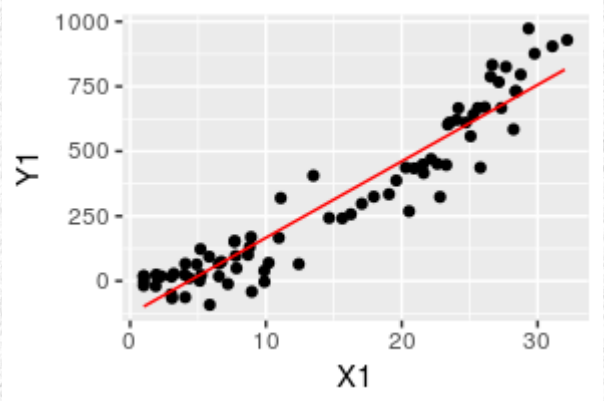
Coefficients table

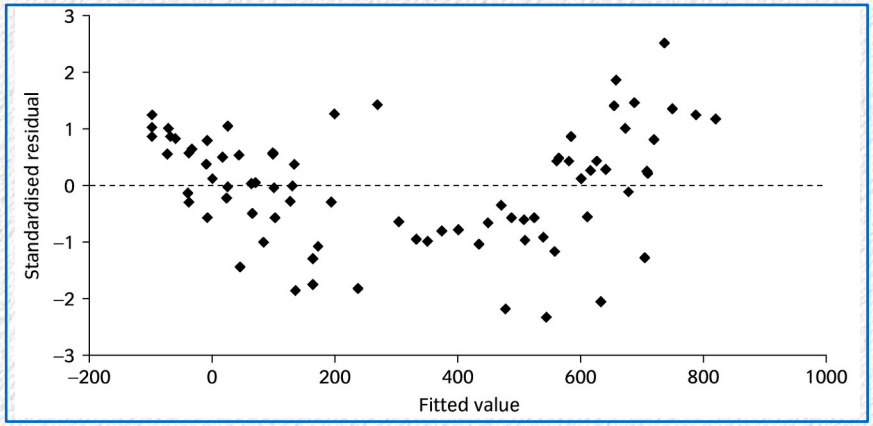
| Term | Coef | SECoef | T | P |
|---|---|---|---|---|
| Constant | −128.08 | 19.40 | −6.60 | 0.000 |
| X1 | 29.473 | 1.096 | 26.89 | 0.000 |

# Linear – poor fit

*Not complicated enough!*

*Add a quadratic term to see if GLM assumptions are met*



General Linear Model

Word equation: Y1 = X1 + X1 * X1

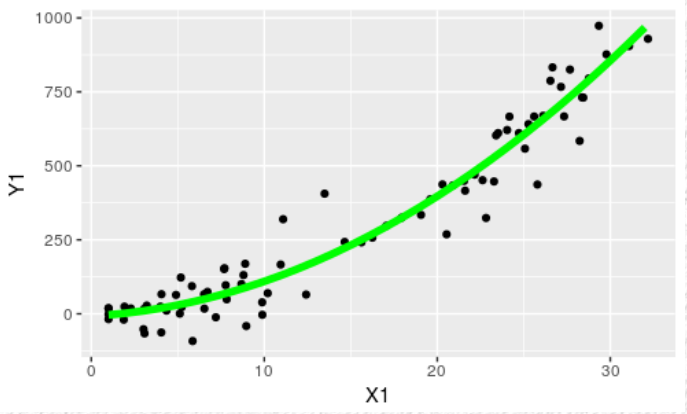
X1 is continuous

Analysis of variance table for Y, using Adjusted SS for tests

| Source | DF | Seq SS | Adj SS | Adj MS | F | P |
|---|---|---|---|---|---|---|
| X1 | 1 | 6 663 021 | 3 597 | 3 597 | 0.60 | 0.442 |
| X1 * X1 | 1 | 256 148 | 256 148 | 256 148 | 42.62 | 0.000 |
| Error | 77 | 462 798 | 462 798 | 6 010 | | |
| Total | 79 | 7 381 967 | | | | |

Coefficients table

| Term | Coef | SECoef | T | P |
|---|---|---|---|---|
| Constant | −7.62 | 24.21 | −0.31 | 0.754 |
| X1 | 3.189 | 4.122 | 0.77 | 0.442 |
| X1*X1 | 0.8525 | 0.1306 | 6.53 | 0.000 |

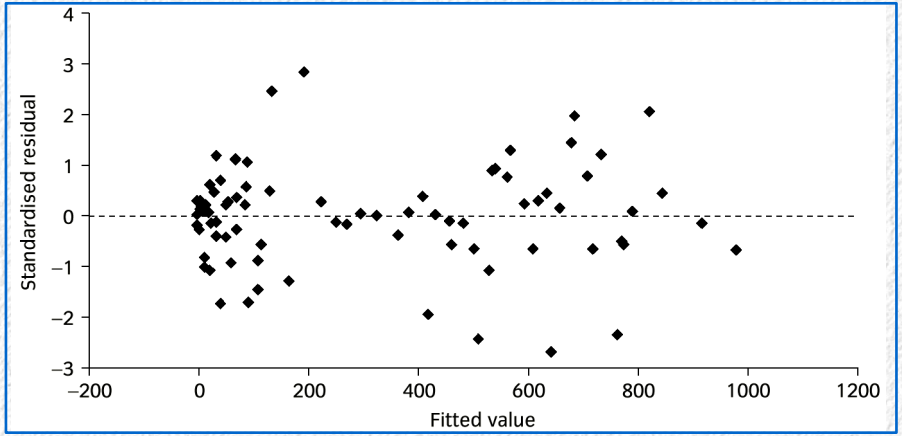



# Including a quadratic term meets GLM assumptions

$$Y1 = -7.62 + 3.189*X1 + 0.825*X1^2$$

*This is complex enough to meet assumptions*

*Should we include a cubic term? How would we know if that makes the model too complex?*



# Using adjusted $R^2$ to pick a model that balances bias and variance

*Adjusted $R^2$ helps us select the best model when predictors are not orthogonal*

$$R^2 = \frac{\text{Total SS} - \text{Residual SS}}{\text{Total SS}}$$

*Multiple $R^2$ always increases with every additional x-variable*

$$R^2_{\text{adj}} = \frac{\text{Total MS} - \text{Residual MS}}{\text{Total MS}}$$

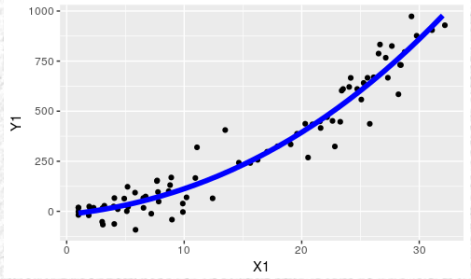
*Adjusted $R^2$ increases when an added variable explains enough variation to compensate for reduced residual d.f.*

*Adding a poor predictor (with $F < 1$) can decrease $R^2_{adj}$*

*Thus, selecting the model with the highest adjusted $R^2$ balances bias and variance*

# Should we include a cubic?

$$Y1 = -15.75 + 6.179*X1 + 0.6169*X1^2 + 0.00500*X1^3$$




General Linear Model

Word equation: Y1 = X1 + X1 * X1 + X1 * X1 * X1

X1 is continuous

Analysis of variance table for Y, using Adjusted SS for tests

| Source | DF | Seq SS | Adj SS | Adj MS | F | P |
|---|---|---|---|---|---|---|
| X1 | 1 | 6 663 021 | 2 505 | 2 505 | 0.41 | 0.523 |
| X1 * X1 | 1 | 256 148 | 4 763 | 4 763 | 0.78 | 0.379 |
| X1 * X1 * X1 | 1 | 720 | 720 | 720 | 0.12 | 0.732 |
| Error | 76 | 462 078 | 462 078 | 6 080 | | |
| Total | 79 | 7 381 967 | | | | |

Coefficients table

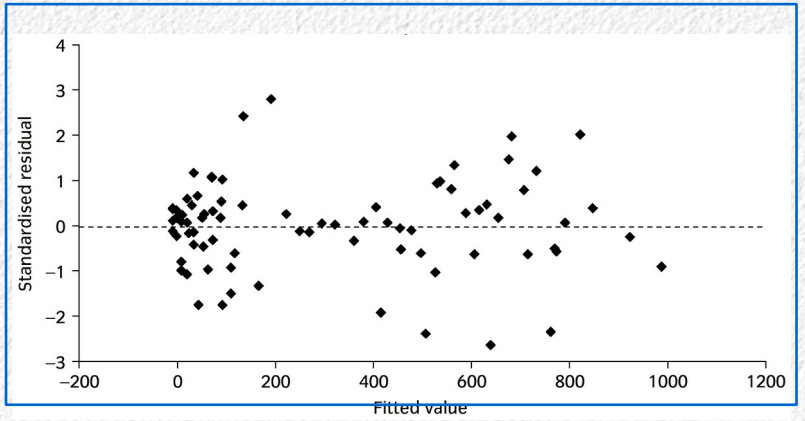
| Term | Coef | SECoef | T | P |
|---|---|---|---|---|
| Constant | −15.75 | 33.92 | −0.46 | 0.644 |
| X1 | 6.179 | 9.625 | 0.64 | 0.523 |
| X1 * X1 | 0.6169 | 0.6971 | 0.89 | 0.379 |
| X1 * X1 * X1 | 0.00500 | 0.01452 | 0.34 | 0.732 |

*Increases the SE's on the coefficients*

*Only tiny increase in $R^2$ – from 0.9373 to 0.9374*

*Adjusted $R^2$ declines slightly from 0.9357 to 0.9349*

*So, no – don't include the cubic*



# Second issue – which set of predictors is best?

- Goal is to build a model that best explains variation in the predictor
- Meeting assumptions is necessary, but not sufficient
  - May meet GLM assumptions with model that has a very low $R^2$
  - May meet GLM assumptions with more than one model
- Can use adjusted $R^2$ as a criterion to compare alternative models, pick the one with the highest adjusted $R^2$

# Example: modeling systolic blood pressure

- Two measures are taken for blood pressure: systolic (heart contraction) and diastolic (rebound of arterial walls)
- Measures of systolic blood pressure for 39 men who had migrated from living at high elevation to low elevation in Peru
- Predictors recorded are:
  - Years (since migration), age
  - Weight, height
  - Chin (skin fold thickness), forearm (skin fold), calf (skin fold)
  - Pulse
- What set of predictors gives the best model of systolic blood pressure?

# Every predictor included – good model?

General Linear Model

Word equation: SYSTOL = YEARS + WEIGHT + AGE + HEIGHT + CHIN + FOREARM + CALF + PULSE

YEARS, WEIGHT, AGE, HEIGHT, CHIN, FOREARM, CALF and PULSE are all continuous.

Analysis of variance for SYSTOL, using Adjusted SS for tests

| Source | DF | Seq SS | Adj SS | Adj MS | F | P |
|---|---|---|---|---|---|---|
| YEARS | 1 | 50.0 | 697.6 | 697.6 | 6.41 | 0.017 |
| WEIGHT | 1 | 2698.3 | 2201.7 | 2201.7 | 20.22 | 0.000 |
| AGE | 1 | 27.9 | 97.4 | 97.4 | 0.89 | 0.352 |
| HEIGHT | 1 | 61.4 | 263.6 | 263.6 | 2.42 | 0.130 |
| CHIN | 1 | 366.9 | 249.3 | 249.3 | 2.29 | 0.141 |
| FOREARM | 1 | 42.7 | 59.2 | 59.2 | 0.54 | 0.467 |
| CALF | 1 | 14.7 | 16.2 | 16.2 | 0.15 | 0.703 |
| PULSE | 1 | 3.0 | 3.0 | 3.0 | 0.03 | 0.870 |
| Error | 30 | 3266.7 | 3266.7 | 108.9 | | |
| Total | 38 | 6531.4 | | | | |

$R^2 = 50.0\%$ $R^2(\text{adj}) = 36.6\%$

*Maybe, can we do better?*

# Poor predictors decrease adjusted $R^2$, increase standard errors of coefficients

*Adjusted $R^2$ is better for simpler model (yellow)*

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.3191    15.8184   3.181  0.00302 **
YEARS        -0.5718     0.1879  -3.043  0.00436 **
WEIGHT        1.3541     0.2672   5.067 1.22e-05 ***

Residual standard error: 10.25 on 36 degrees of freedom
Multiple R-squared:  0.4208,    Adjusted R-squared:  0.3886
F-statistic: 13.08 on 2 and 36 DF,  p-value: 5.385e-05
```

*Standard error for years (blue) and weight (orange) are smaller for simpler model*

*$R^2$ higher for complex model, but would be true even if we generated random data*

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 127.40514   58.72121   2.170   0.0381 *
YEARS        -0.57359    0.22662  -2.531   0.0168 *
WEIGHT        2.11448    0.47024   4.497 9.61e-05 ***
AGE          -0.27493    0.29068  -0.946   0.3518
HEIGHT       -0.06727    0.04323  -1.556   0.1302
CHIN         -1.33839    0.88460  -1.513   0.1407
FOREARM      -1.06039    1.43822  -0.737   0.4667
CALF          0.24467    0.63499   0.385   0.7027
PULSE         0.03357    0.20388   0.165   0.8703

Residual standard error: 10.44 on 30 degrees of freedom
Multiple R-squared:  0.4998,  Adjusted R-squared:  0.3665
F-statistic: 3.748 on 8 and 30 DF,  p-value: 0.003783
```

*So, the simpler model is preferred for these data*

# Problem: groups of variables

- Sometimes effects of one variable depend on inclusion of another
  - Variables may be confounded – adding or removing one while others are in may not improve adjusted $R^2$
  - The effect of one variable may be strong, but only after another is included – adding the variable alone may not improve adjusted $R^2$, but adding both together would
- Solution: can add or remove groups of variables at once
  - Can test for statistical significance of groups of variables by combining the terms

General Linear Model

Word equation: SYSTOL = YEARS + WEIGHT + AGE + HEIGHT + CHIN + FOREARM + CALF + PULSE

YEARS, WEIGHT, AGE, HEIGHT, CHIN, FOREARM, CALF and PULSE are all continuous.

Analysis of variance for SYSTOL, using Adjusted SS for tests

| Source | DF | Seq SS | Adj SS | Adj MS | F | P |
|---|---|---|---|---|---|---|
| YEARS | 1 | 50.0 | 697.6 | 697.6 | 6.41 | 0.017 |
| WEIGHT | 1 | 2698.3 | 2201.7 | 2201.7 | 20.22 | 0.000 |
| AGE | 1 | 27.9 | 97.4 | 97.4 | 0.89 | 0.352 |
| HEIGHT | 1 | 61.4 | 263.6 | 263.6 | 2.42 | 0.130 |
| CHIN | 1 | 366.9 | 249.3 | 249.3 | 2.29 | 0.141 |
| FOREARM | 1 | 42.7 | 59.2 | 59.2 | 0.54 | 0.467 |
| CALF | 1 | 14.7 | 16.2 | 16.2 | 0.15 | 0.703 |
| PULSE | 1 | 3.0 | 3.0 | 3.0 | 0.03 | 0.870 |
| Error | 30 | 3266.7 | 3266.7 | 108.9 | | |
| Total | 38 | 6531.4 | | | | |

$R^2 = 50.0\%$ $R^2(\text{adj}) = 36.6\%$

*Skin fold thickness still not significant when all three are grouped – lack of significance is not due to confounding between them*

# Combining terms – testing related variables as a group

| | DF | SeqSS |
|---|---|---|
| | 1 | 366.9 |
| | 1 | 42.7 |
| + | 1 | 14.7 |
| | 3 | 424.3 |

MS=424.3/3=141.4

F = 141.4/108.9 = 1.30

on 3 and 30 DF

P-value: p=0.293

# Automating the search – stepwise regression

- Meant to help find the best from a large number of possible models
- Stepwise regression = automated model construction based on a set of rules
  - An initial model is selected, then terms are added or dropped one at a time
  - If a variable is dropped and fit goes down substantially, the variable is put back in
  - If a variable is added and it does not contribute to an increase in fit, it is omitted
- This process is repeated until no further improvements are found

*Worked example of blood pressure data on course web page...*

# Criticisms of stepwise procedures

- Machine intelligence is not as good as real intelligence
- Groups of variables may need to be entered or removed
- Investigator's choice of forward vs. backward selection
  - Forward selection = starting simple, adding variables each step
  - Backward selection = starting with all variables included, removing variables each step

  ...sometimes arrive at different models, so which to use?
- If the final model depends on judgment, better to make the decisions yourself

# Models are hypotheses about your data

- The model you build is a statement of a hypothesis about the structure in the response variable
  - Including a predictor in a model is a hypothesis that it affects the response
  - Omitting a predictor is a hypothesis that the response is independent of it
  - Including an interaction is a hypothesis that the effect of one predictor depends on the level of another
  - Including a polynomial term (squared, cubic, etc.), or log-transforming predictor or response hypothesizes a non-linear relationship
  - The levels used in a categorical variable is a hypothesis that there will be differences on average between those groups, and only between those groups
- Adjusted $R^2$ is then a measure of which hypothesis is best supported by the data
- We will use model selection to test hypotheses for the rest of the semester

# In summary...

- Model selection allows us to seek the best representation for the data in hand
- Important principles:
  - Balancing bias against variance ($R^2$ against standard errors of estimates)
  - Economy of variables: use adjusted $R^2$ to avoid making models too complex
  - Models need to be complex enough to meet GLM assumptions
  - Avoiding multiplicity of p-values: simplify as much as possible, only fit reasonable models
- Orthogonal designs make model choice simpler
- Stepwise procedures automate model choice, but the final models depend on analysts choices – given this, it's usually better to build and evaluate your own

# What is the model?