Week 8

Hour 1: More on polynomial fits. The AIC

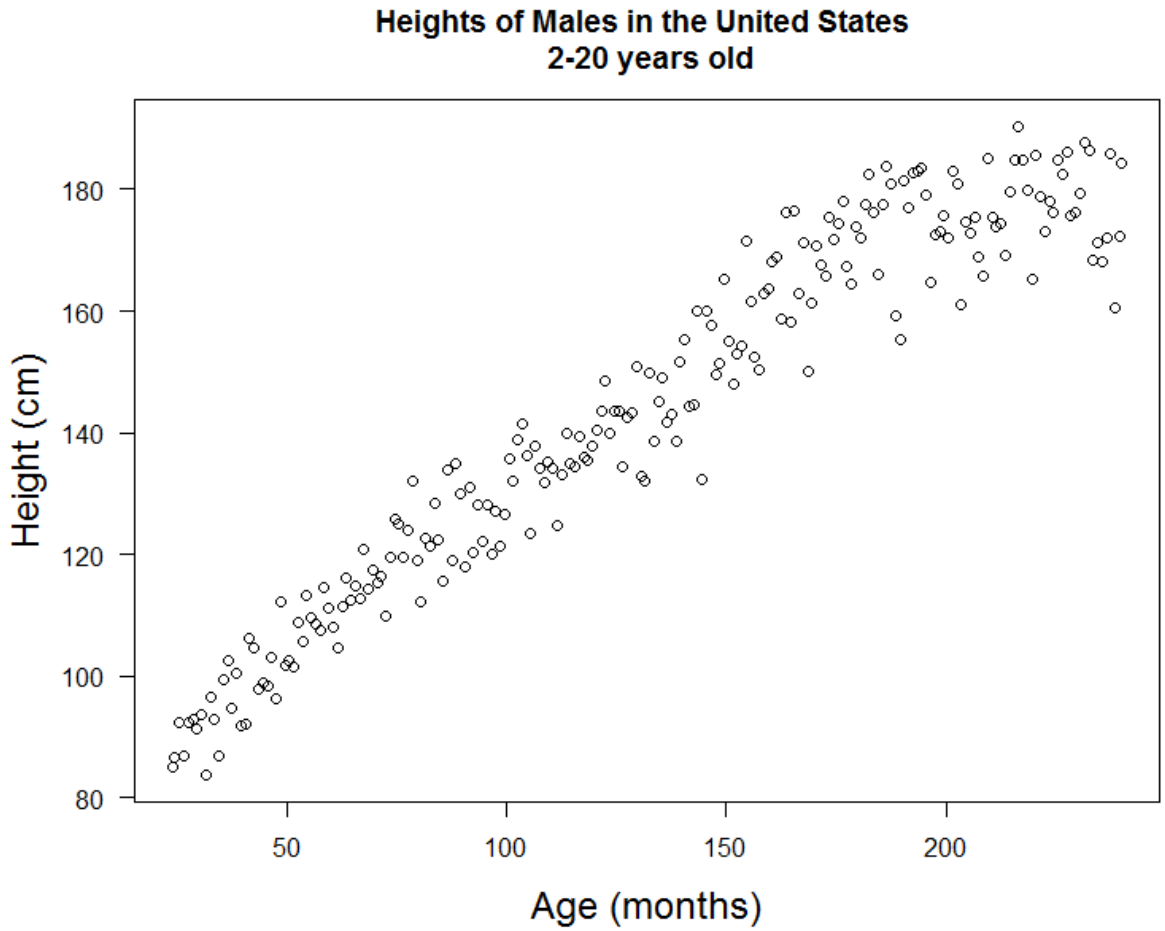
Hour 2: Dummy Variables – what are they? An NHL Example

Hour 3: Interactions. The stepwise method.

Human growth example for polynomial fits.

Consider the dataset USheights.csv, which is a sample of heights at different months of people aged 24 to 240 months, generated from the growth curves given at the US Centre for Disease Control.

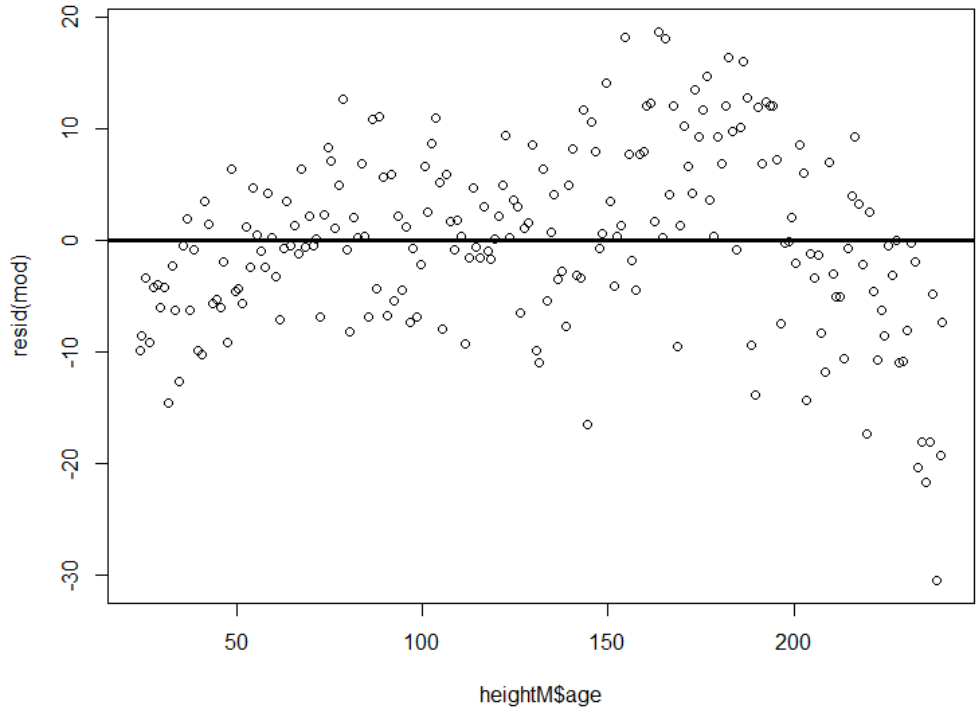
They reach a maximum after 240 months, so we are only considering this range. The heights of males look like this:



We will start with a simple regression as a comparison point.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 84.183781   1.270847   66.24   <2e-16 ***
age          0.448160   0.008691   51.57   <2e-16 ***
```

```
Multiple R-squared:  0.9249
```



Now we'll try to include a squared term.

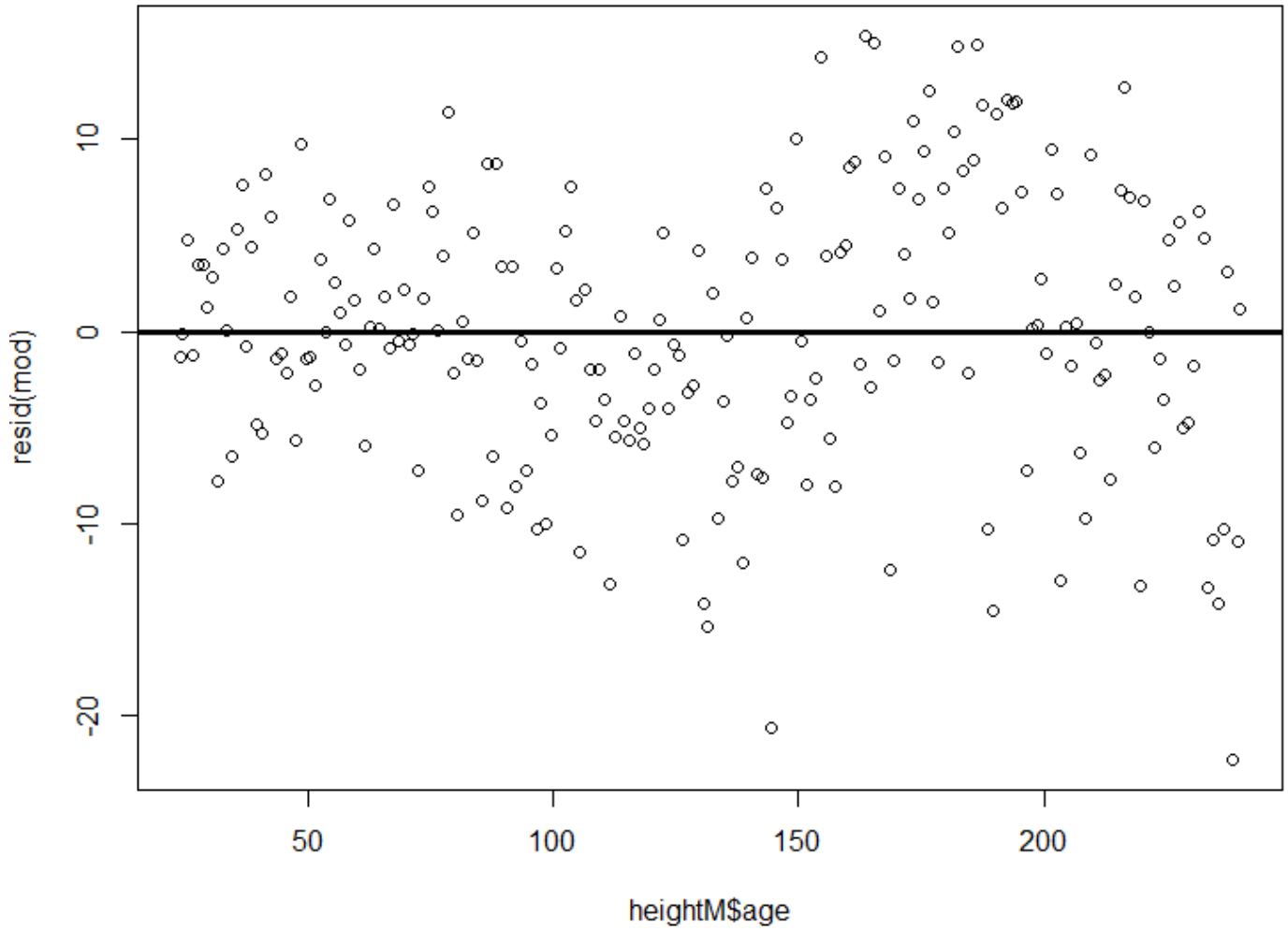
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 69.3119313  2.1346194  32.470  < 2e-16 ***
age          0.7397488  0.0365165  20.258  < 2e-16 ***
I(age^2)    -0.0011045  0.0001353  -8.164 2.72e-14 ***

Multiple R-squared:  0.9427
```

The r-squared has improved from 0.9249 to 0.9427

The age-squared term is showing up as highly significant, these are both good signs.

Have the residuals improved?



Mostly. There's still a trend, but not a huge curve.

Let's see if a square-root term does a better job than a squared term.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.83755    6.50529   6.124 4.29e-09 ***
age           0.03362    0.06037   0.557    0.578
I(sqrt(age))  8.92146    1.28804   6.926 4.93e-11 ***

Multiple R-squared:  0.9386,
```

Including a square root age instead of age-squared still improves the variance explained, but not as much.

Since both the square root term and the squared term each cost a single degree of freedom, we'll take the one with the better r-squared.

However, the variance inflation factors for the age-squared model are large.

```
> vif(mod)
      age  I(age^2)
23.02153 23.02153
```

Previously, we had said that a VIF of 5 or more is a potential problem because...

- It increased the uncertainty in the coefficients.

- It indicated that two or more of the explanatory variables were explaining the same variance.

Comparing the simple model,

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 84.183781   1.270847   66.24   <2e-16 ***
age          0.448160   0.008691   51.57   <2e-16 ***
```

and the model with age squared.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 69.3119313  2.1346194  32.470  < 2e-16 ***
age          0.7397488  0.0365165  20.258  < 2e-16 ***
I(age^2)    -0.0011045  0.0001353  -8.164 2.72e-14 ***
```

The standard error for age HAS increased from 0.009 to 0.037. However, we still know it's highly significant.

Also, the age term doesn't have a real-world interpretation.

The slope value 0.7397 means

'Height increases by 0.7397 cm per month, holding age-squared constant'

...but there is no way to change age and hold age-squared constant.

That's part of the cost of using polynomial models, is that the coefficients are sometimes impossible to interpret in real-world terms.

However, VIF does not harm prediction value. So if we were trying to predict how tall someone will be at different ages, then the regression equation:

$$\text{Height} = 69.31 + 0.7397*\text{Age} - 0.0011*\text{Age}^2 + \text{error}$$

is perfectly adequate, especially since the higher r-squared means the errors are smaller, and because there less of a pattern left in the residuals.

Inflation isn't always a problem

AIC

AIC stands for ***Akaike information criterion***

It's measure of two features of a model:

- **Goodness of fit**: How well does the model predict the observed data.

- **Complexity**: The number of terms in the model.

Compare this to the proportion of variance explained, R^2.

R^2 only measures goodness of fit.

R^2 always increases when you add variables to the model. Even when those variables explain no new variance, or have nothing to do with the model, R^2 will at worst stay the same.

It is useful for comparing models that have the same number of terms.

AIC can be used to compare many different models, as long as they have the same response variable and observations.

AIC is calculated by...

$$\mathrm{AIC} = 2k - 2\ln(L)$$

where k is the number of ***parameters*** in the model

(or the number of ***degrees of freedom*** being used up)

ln(L) is the 'log likelihood', which is a measure of how well the model fits the data.

$$\mathrm{AIC} = 2k - 2\ln(L)$$

For AIC, lower numbers are better.

So the 2k part is the 'penalty term' that imposes a penalty of 2 points for every variable included in the model.

-2ln(L), or 'negative log likelihood' is like R^2 in the sense that it can only go one direction when terms are added: down.

$$\mathrm{AIC} = 2k - 2\ln(L)$$

A term is worth including in the model if makes the AIC lower.

In other words, if it explains enough of the variation to make -2ln(L) decrease by more than the 'cost' of including it,

The cost is 2 times the df.

For a regression term, that's just 2.

For an ANOVA grouping variable that 2*(Ngroups – 1)

Usually, if a variable has statistical significance (at alpha = 0.05), then the AIC is improved by having that term in the model.

However, because of co-linearity, sometimes that variable is 'stealing' the significance from some other term. The AIC doesn't care which terms are significant, it just looks at how well the model fits as a whole.

Consider the USheights data-set again.

The model with just age has an AIC of 1533.

```
> mod = lm(height ~ age, data=heightM)
> AIC(mod)
[1] 1533.302
```

The model with age and age-squared has an

AIC of ***1476***, using 3 df.

```
> mod = lm(height ~ age + I(age^2), data=heightM)
> AIC(mod)
[1] 1476.43
```

The model with age-squared has an AIC that is 57 points better than the simpler model. That's very strong evidence that including age squared is an improvement.

For regression and ANOVA, an improvement in AIC of ***2 or more points*** is considered statistically significant at alpha = 0.05.

An improvement of ***3 or more points*** is significant at alpha = 0.001.

Consider also the gapminder birth rate model

When we include all six variables, the AIC is 192.2

```
> mod = lm(birth_rate ~ agri_in_gdp + GINI + GDPpercap +
           HDI + health_spending + female_work, data=gapminder)
> AIC(mod)
[1] 192.2269
```

We removed the health_spending variable because of its high VIF. When we do that, the AIC also improves

```
> mod = lm(birth_rate ~ agri_in_gdp + GINI + GDPpercap +
           HDI + female_work, data=gapminder)

> AIC(mod)
[1] 191.7546
```

But when we also remove HDI, which turned out to be a bad idea before, the AIC gets worse. A lot worse.

```
> mod = lm(birth_rate ~ agri_in_gdp + GINI + GDPpercap +
           female_work, data=gapminder)
> AIC(mod)
[1] 215.443
```

So in a way, this single measure can do a lot of the judgement work for us.

Have you ever thought there could be more to being a statistical model than being really, really, ridiculously well-fitting?

BIC

BIC stands for ***Bayesian Information Criterion***.

It's also called the Schwarz Criterion (SC).

It behaves very similarly to the AIC, but imposes a larger penalty term for complexity.

Also, like the AIC, a smaller value is better, and 2 points is significantly better.

The practical difference between the BIC and the AIC is that the BIC favours simpler models.

When given a set of candidate models, the model with the best BIC will sometimes be simpler than the model with the best AIC.

(Technically this is only true with there are more than 8 observations in dataset, which is almost always)

BIC is calculated by...

$$BIC = (ln(n) \times k) - 2log(L)$$

where n is the ***number of observations***, also called the sample size.

As in the Akaike Information Criteria (AIC), k stands for the number of parameters, and -2log(L) is the negative log likelihood.

$$BIC = (ln(n) \times k) - 2log(L)$$

This means that the penalty term per parameter gets larger more observations are taken.

The idea behind the BIC is that it is easier to find significance in variables that are unimportant when n is large.

$$BIC = (ln(n) \times k) - 2log(L)$$

Recall from Assignment 2 that a weak correlation from n=50 points can be found significant even though a stronger correlation from n=10 points cannot.

BIC, compared to AIC serves to counter this effect of increasing n by increasing the penalty accordingly.

Consider the USheights data-set one more time

The model with just age has a BIC of 1543.

```
> mod = lm(height ~ age, data=heightM)
> BIC(mod)
[1] 1543.455
```

The model with age and age-squared has an

BIC of ***1489***, using 3 df.

```
> mod = lm(height ~ age + I(age^2), data=heightM)
> BIC(mod)
[1] 1489.968
```

Now consider the gapminder birth rate model again.

The original six-variable model has a BIC of 204.20

```
> mod = lm(birth_rate ~ agri_in_gdp + GINI + GDPpercap +
+ HDI + health_spending + female_work, data=gapminder)
> BIC(mod)
[1] 204.199
```

The five-variable model has an improved BIC, but the improvement is larger because of the larger penalty.

```
> mod = lm(birth_rate ~ agri_in_gdp + GINI + GDPpercap +
+ HDI + female_work, data=gapminder)
> BIC(mod)
[1] 202.2301
```

As with the AIC, removing both HDI and health_spending produces a model that is much worse.

In this case, 20 points worse.

```
> mod = lm(birth_rate ~ agri_in_gdp + GINI + GDPpercap +
+ female_work, data=gapminder)
> BIC(mod)
[1] 224.4221
```

AIC and BIC with ANOVA

Recall the chickwts dataset from Assignment 2, and recall that feed had six groups.

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

We found that feed was a significant grouping variable.

The AIC and BIC for this one-way ANOVA are:

```
mod = lm(weight ~ feed, data=Q3)
```

```
> AIC(mod)
[1] 777.8748
```

```
> BIC(mod)
[1] 793.7135
```

But if we remove the feed variable and just take a model without any explanatory variables.

```
> mod = lm(weight ~ 1, data=Q3)
> anova(mod)
Analysis of Variance Table

Response: weight
           Df Sum Sq Mean Sq F value Pr(>F)
Residuals 70 426685  6095.5
```

The AIC and BIC both get worse.

```
> AIC(mod)
[1] 823.2689
```

```
> BIC(mod)
[1] 827.7943
```

This should make sense, we have removed a grouping variable that was improving the model fit.

Next hour: Dummy Variables!