Testing assumptions of linear regression

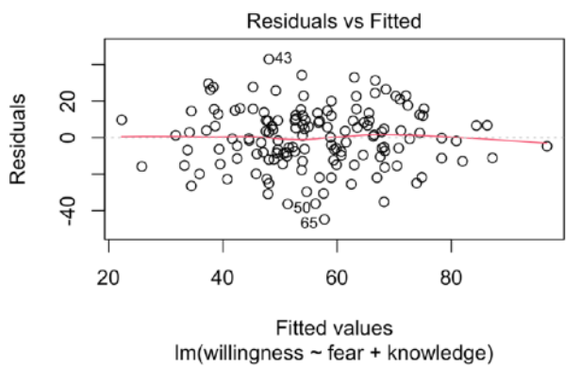
1

Linearity: test by looking at the residuals on each fitted value. If even across all, probably linear

```
> library(car)
> residualPlots(model)
```

or

```
> plot(model,which=1)
```



2

Normality of the residuals: test by extracting the residuals and then using standard tests of normality

```
> resids <- rstandard(model)
```

QQ plot

```
> qqnorm(resids)
```

Shapiro-Wilk

```
> shapiro.test(resids)
```

3

Identify high-influence points: calculate Cook's Distance, which captures both outlierness and leverage. High-influence points are high on both

Measures outlier-ness

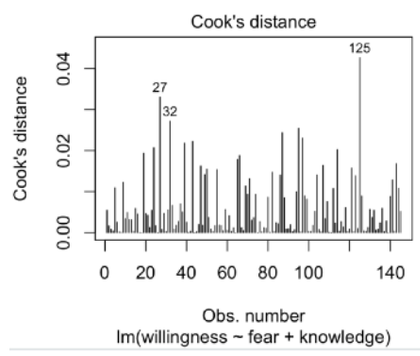
$$D_i = \frac{\epsilon_i^{*2}}{K+1} \times \frac{h_i}{1-h_i}$$

Measures leverage

```
> cooks.distance(model)
```

or

```
> plot(model,which=4)
```



Many rules of thumb. Common one is if Cook's distance is greater than around 0.1 or 1. In this subject we'll be more conservative: if > $2k/N$, where k=# of coefficients, you might have an issue

4

Identify collinearity: variables that are highly collinear contribute similar information, and can make the regression not robust. Quantify with Variance Inflation Factor, which captures how badly the correlation is messing up your estimate of the coefficients

$$\mathrm{VIF}_k = \frac{1}{1-R^2_{(-k)}}$$

Variance accounted for by the model where variable *k* is the outcome and the others are predictors

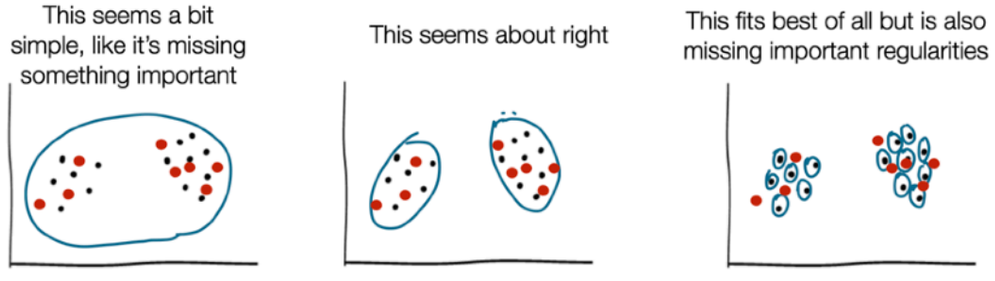
```
> vif(model)
```

VIF = 1 is great!
Much larger than 2 or 3.. possible problem?
Don't use VIF on models with interactions!

Model selection: choosing between models

1

More complex models always reduce variance (or at least don't increase it) but you don't always just want those, otherwise you'll **overfit**



2

Lots of ways to penalise model complexity, usually by penalising extra parameters

AIC: Akaike Information Criterion

$$\mathrm{AIC} = \frac{\mathrm{SS}_{res}{}^{2}}{\hat{\sigma}} + 2K$$

How big the residuals are relative to your variance

Two times the # of parameters in your model

BIC: Bayesian Information Criterion

$$\mathrm{BIC} = k\ln(n) - 2\ln(\widehat{L}).$$

The number of parameters times the natural log of the sample size

Essentially a measure of how well the model fits (similar to the residual term in AIC)

3

Using AIC() and BIC() in R: Give them the list of model objects you want to compare. Key is that the best has the *lowest* AIC or BIC

```
> AIC(model1,model2,model3)
                df   AIC
model1           3 765.32
model2           2 678.54
model3           4 702.45

> BIC(modelA,modelB,modelC)
                df   BIC
modelA           4 657.09
modelB           3 694.32
modelC           4 633.36
```

4

Resist the temptation to just include every single model. Complexity penalisations are art as much as science, and there's no point in having a model that is so complex you can't interpret the parameters. Compare the models that are theoretically interesting and interpretable.