# Prediction in regression

ST552 Lecture 11

Charlotte Wickham

2019-02-01

# Next week

Midterm in class on Friday. Posted under today's date:

- Study guide
- Previous year's midterm (and solution)

Homework 4 has a due date two weeks away (you decide if you want to do it now or later)

Next week:

- Weds lecture: review. Bring your questions.
- Weds lab: Trevor will go over a relevant comp exam question (you might like to look it over beforehand)

# Prediction

We've built a model:

$$y = X\beta + \epsilon$$

Now given a new vector of values of the explanatories $x_0$ we can predict the response

$$\hat{y}_0 = x_0^T \hat{\beta}$$

But what is the uncertainty in this prediction?

Two kinds:

- prediction of the mean response
- prediction of a future observation

# Faraway example

Suppose we have bulit a regression model that predicts the rental price of houses in a given area based on predictors such as the number of bedrooms and closeness to a major highway.

Two kinds of predictions:

- Prediction of a future value
- Prediction of the mean response

# Prediction of a future value

Suppose a specific house comes on the market with characteristics $x_0$. It's rental price will be $x_0^T \beta + \epsilon$.

Since, $\mathrm{E}(\epsilon) = 0$ our predicted price will be $x_0^T \hat{\beta}$, but in assessing the variance of this prediction, we must include an estimate of $\epsilon$.

*Our uncertainty comes from our uncertainty in our estimates, as well as the variability of the response about its mean*

# Prediction of the mean repsonse

Suppose we ask the question – "What would a house with characteristics $x_0$ rent for on average?"

This price is $x_0^T \beta$ and is again predicted by $x_0^T \hat{\beta}$ but now only variance in $\hat{\beta}$ needs to be taken into account.

*Our uncertianty only comes from our uncertainty in our estimates*

# Leads to two types of interval

$$\text{Var}\left(x_0^T \hat{\beta}\right) = \sigma^2 x_0^T (X^T X)^{-1} x_0$$

Assuming future $\epsilon$ is independent of $\hat{\beta}$ a **prediction interval** for a future response is:

$$\hat{y_0} \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0}$$

A **confidence interval** for the mean response is:

$$\hat{y_0} \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0}$$

which will always be narrower.

# Work through Faraway's example

> *Normally, we would start with an exploratory analysis of the data and a detailed consideration of what model to use but let's be rash and just fit a model and start predicting.*

Find Rmarkdown in rstudio.cloud or get at:
stat552.cwick.co.nz/lecture/11-faraway-fat.Rmd

We are interested in predicting body fat (%) as a function of physical measurements (e.g. weight, height, circumference of hip, etc.)

(*Different data to lab, remember there we were predicting weight*)

# Your Turn #1

```
?fat
```

Take a quick read through of the documentation on this dataset.

In context of the data (discuss with your neighbours):

- What would a confidence interval on the mean response tell us? When might it be useful?
- What would a prediction interval on a response tell us? When might it be useful?

# Your Turn #2

Go through the code. Discuss each step:

- what is happening conceptually?
- what is the code doing?
- are their other ways to do it?

There are **three questions** for you in the code. Work with your neighbours to answer them.

# Notes

- Prediction intervals become wider the further we are from observed values of the predictors.
- These intervals depend on the model being correctly specified. In practice, you never know the true model. We do our best to specify a good model but there is always uncertainty in the form of the model.
  This *model uncertainty* is not reflected in these intervals. We take into account *parameter uncertainty*, but *model uncertainty* is harder to quantify.

# What can go wrong with predictions?

1. Bad model
2. Quantitative extrapolation (explanatory values beyond those observed)
3. Qualitative extrapolation (situations beyond those which generated the data, i.e. extrapolation to a different population, e.g. females)
4. Overconfidence due to overfitting
5. Black swans (very unusual events that don't occur in sample, but do occasionally occur)