Regression analysis: a primer

Data to decisions

The universal mathematics of linear regressions

- Take two random variables X and Y with finite variance

- Then it is **always** possible to write:

$$Y = a + bX + \varepsilon$$

where: $E(\varepsilon) = 0$ and $COV(X, \varepsilon) = 0$

- The model $a + bX$ is the best possible linear relationship between Y and X in the sense that $VAR(\varepsilon)$ is the lowest it can possibly be

R-squared ("goodness of fit")

- Since $COV(X,\varepsilon) = 0$,

$$VAR(Y) = b^2 VAR(X) + VAR(\varepsilon)$$

- Total Variance = Variance explained by the model + Residual Variance

- Then,

$$R^2 = \frac{b^2 VAR(X)}{VAR(Y)}$$

tells us what fraction of the variance of Y is "explained" by the model

Regressions in statistics

- Assume we get a sample X_i, Y_i of joints observations of/draws from X and Y for $i = 1,2, \dots n$

- We can plot the resulting data with Y on the vertical axis and fit the best possible line through these dots

- This gives us estimates $\hat{a}$ and $\hat{b}$ of a and b

- Furthermore, by the law of large numbers, $\hat{a} \to a$ and $\hat{b} \to b$ as n gets large

- So now if you give me any particular X I can forecast Y as $\hat{a} + \hat{b}X$

- This is the best linear forecast I can make, at least in sample

Confidence and prediction intervals for linear forecasts

- How confident should I be in my forecast?
- After all:
    1. I am uncertain about a and b
    2. I don't know what ε draw I am going to get
- The first issue affects my ability to know $E(Y|X)$
- *Confidence intervals* reflect <u>only</u> that first source of uncertainty
- My ability to forecast the Y value of one specific new observation is also limited by the ε problem
- *Prediction intervals* reflect both sources of uncertainty
- They tend to be very, very large even in pretty good R^2 situations

Significance test

- $H0: b = 0$
- Can $H0$ be rejected?
- Under the assumption that Y is normally distributed, the standard error $\sigma(\hat{b})$ of $\hat{b}$ can be computed…
- … and $\mathrm{T} = \frac{\hat{b}}{\sigma(\hat{b})}$ follows a t-distribution with $n-2$ degrees of freedom
- Basically and with enough data, reject $H0$ if $T > 1.96$ or $T < -1.96$

Classical assumptions

- Classical regression based inference relies on three main assumptions:
  1. The error terms are normally distributed
  2. They are independent of X (*homoscedasticity*)
  3. They are independent from one another

- Errors that satisfy those assumptions are called *spherical*

- If they do then all the tests and confidence intervals we have developed so far are valid

Diagnosis

- The obvious ways to detect issues are to
  1. plot residuals and look at the shape of the distribution
  2. plot residuals against X and look for patterns

- There are formal tests that automate this

Broad remedies

- Play with the model specification (go from Y to $\log Y$ to deal with curvature issues…)
- Look for missing variables
- Understand the pattern in error dependence and use *GLS*

Outliers

- Sometimes your plots will show observations that are way off, that visibly stand alone
- There are tests that detect those
- Two possibilities: *contaminated case* or *rare case*
- In case 1, drop or correct the observation, obviously, but make sure the same contamination does not pollute the rest of your data
- In case 2, you need to model rare case and typical case separately, maybe by mixing models
- Sometimes, (in value-at-risk management, say, or mortgage design) it’s all about rare cases

Multivariate case

- If we add more *explanatory variables*, nothing of importance changes

- Say, $Y = a + b_1X + b_2Z + \varepsilon$

- We can only improve fit by adding variables (but fit is not the goal, more on that in next chapter)

- Now we can test joint hypothesis, like $H0{:}\, b_1 = b_2$, using what's called an F-test, which any stats package can perform for you

- And we can still test the individual significance of each coefficient using t-tests

Forecasting with log transforms

- When $\ln(Y)$ is the dependent variable, the error in logs is minimized
- Negative errors are more penalized than positive errors (asymmetric *loss function*)
- $\exp E(\widehat{\ln(Y)}|X) \leq E(\hat{Y}|X)$, a fact known as Jensen's inequality

- If 1) the model is well specified and 2) errors are spherical, then an unbiased forecast is:

$$\exp E(\widehat{\ln(Y)} + s^2/2\,|X)$$

where s^2 is $Var(\hat{\epsilon})$

- <u>Remark 1</u>: bias is often small
- <u>Remark 2</u>: correction above may do more harm than good when either assumption is badly violated
- <u>Remark 3</u>: prediction intervals are correct under naïve transform, though they can be improved