Math 455 Class 26 March 25

Last time we discussed VIF (Variance Inflation Factors). What are these?

Suppose we do a regression

$X_i \sim$ (all other X_j s).

e.g. $X_1 = \beta_0 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}$

(in Faraway notation with p ~~predictors~~ parameters total)

We get a R^2 for this regression.

With X_i on LHS, call this R_i^2.

ICBST:

$$Var(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1-R_j^2} \right) \frac{1}{\sum_i (x_{ij} - \bar{x}_j)^2}$$

This is the VIF: $(1-R_j^2)^{-1}$

That is if X_j is well-explained by the other predictors, R_j^2 is close to 1, and VIF is large.

Note that we do not actually do the computation of these as on the previous page!

There is a "vif" command in the Faraway package! Example in Script for Ch. 7.

This was part of our "how do we detect collinearity?" discussion. There was a separate question: "what do we do about collinearity?" and we gave several answers. Here are a couple of more which are beyond the scope of this course.

Option 6: Get more data, specifically chosen to alleviate the collinearity problem (Experimental Design)

Option 7: Make up more data, perhaps by adding random errors to the existing data. (Just don't pretend it's real!)

From last time: Change of Scale in Regression (7.2)

We consider a simple regression

$Y =$ weight (in kg) $X =$ height (in meters)

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

We get our regression output:

$\hat{\beta}_0$, $\hat{\beta}_1$, t-stats, F-stat, $\hat{\sigma}^2$, R^2.

Now we change X to be measured in cm.

$$X_{new} = 100 X_{old}.$$

How does the regression output change?

Answer (textbook p. 104)

$\hat{\beta}_0$, t-stats, F-stat, $\hat{\sigma}^2$, R^2 UNCHANGED.

$\hat{\beta}_1$ (and std. error of $\hat{\beta}_1$) do change.

$$\hat{\beta}_{1\,(new)} = \frac{1}{100} \hat{\beta}_{1\,(old)}$$

Why should this be? Think about the equation

$$\boxed{\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X}$$

Apply dimensional analysis: the idea (most common in physics) if LHS is in miles per hour and RHS is in pounds per square inch, something is wrong.

$\hat{Y}$ is in kg. thus $\hat{\beta}_0$ is in kg.

$\hat{\beta}_1$ (old) has units of $\frac{kg}{m}$

X_{old} has units of meters.

X_{new} has units of cm, $X_{new} = 100 X_{old}$.

Thus $\hat{\beta}_{1\,(new)}$ must have units of $\frac{kg}{cm}$.

We must have $\hat{\beta}_{1\,(new)} \cdot X_{new} = \hat{\beta}_{1\,(old)} X_{old}$.

Thus $\hat{\beta}_{1\,(new)} = \frac{1}{100} \hat{\beta}_{1\,(old)}$ so that

$$\hat{\beta}_{1\,(new)} X_{new} = \hat{\beta}_{1\,(old)} \cdot \frac{1}{100} \cdot 100 \cdot X_{old} = \hat{\beta}_{1\,(old)} X_{old}.$$

Intuition: we can't change the statistical quality of linear regression with linear (or affine) changes of variable.

Thus our statistics and p-values do not change.

Things related to units of measurement do have to change.

Note that this does NOT apply to NON-linear changes of variable, e.g. $X \longmapsto X^2$, $X \mapsto \log X$.

Now we consider $Y_{new} = 1000\, Y_{old}$.

Y_{old} measured in kg. Y_{new} measured in g.

Suppose we run the regression

$$Y_{new} = \beta_0 + \beta_1 X_{old} + \varepsilon.$$

How do all these quantities change vs. the original regression

$$Y_{old} = \beta_0 + \beta_1 X_{old} + \varepsilon$$

Again we consider the equation

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

$$\hat{Y}_{new} = 1000\,\hat{Y}_{old}.$$

thus $\hat{\beta}_0\,(new) = 1000\,\hat{\beta}_0\,(old)$

and $\hat{\beta}_1\,(new) = 1000\,\hat{\beta}_1\,(old)$.

the same is true for the s.d. of $\hat{\beta}_0, \hat{\beta}_1$.

and $\hat{\sigma}$ (since std. dev has the same units as the original measurement.)

But the t-stats, F-stat, R^2, p-values are unchanged.