Unit 5 – Regression and Correlation
Practice Problems (3 of 3)

Due: Thursday October 31, 2024
Last date to submit late for credit (-20 points): Thursday November 7, 2024

Before you begin. Download from the course website
hersdata_small.xlsx

Description of Dataset

Source
Hulley et al (1998) Randomized trial of estrogen plus progestin for secondary prevention of heart disease in postmenopausal women. The Heart and Estrogen/progestin Replacement Study. *Journal of the American Medical Association*, **280**(7), 605-613

The Heart and Estrogen/progestin Replacement Study (HERS) was a randomized clinical trial of hormone therapy (estrogen plus progestin) for the reduction of cardiovascular disease risk in post-menopausal women with established coronary disease. Study participants were n=2,763 women who were: (1) post-menopausal (2) with coronary disease; and (3) with an intact uterus.

The data set for this homework is a random sample of n=1000 from the data set **`hersdata_small.xlsx`**. The following variables are considered:

Data dictionary/Codebook *(Partial)*

Variable	Label	Type	Codings
age	Age, years	numeric	Continuous, range, [45:79]
BMI	Body Mass index (kg/m^2)	numeric	Continuous, range, [15.21:54.13]
glucose	Fasting glucose (mg/dL)	numeric	Continuous, range, [29:298]
LDL	LDL cholesterol (mg/dL)	numeric	Continuous, range, [44.4:393.4]
drinkany	Any current alcohol use	numeric	1 = yes 0 = no
exercise	Exercise at least 3x/week	numeric	1 = yes 0 = no
HT	Randomization	numeric	1 = hormone therapy 0 = placebo
physact	Comparative ("compared to other women your age") physical activity	Numeric	1 = much less active 2 = somewhat less active 3 = about as active 4 = somewhat more active 5 = much more active
statins	Statin use	Numeric	1 = yes 0 = no
diabetes	Diabetes	Numeric	1 = yes 0 = no

The exercises in this assignment give you practice performing regression diagnostics.
They are ***not*** an illustration of an entire regression analysis, beginning with data exploration followed by a series of model estimation followed by diagnostics.

#1
Fit. Using as your dependent variable Y=glucose, fit the following 3 predictor model:

$$glucose = \beta_0 + \beta_1 \cdot age + \beta_2 \cdot BMI + \beta_3 \cdot drinkany$$

Check: You should get the following prediction equation.

$$Predicted\ glucose = 80.07 + 0.056 \cdot age + 0.484 \cdot BMI - 0.388 \cdot drinkany$$

#2.
Linearity of Y in the predictors. Normal theory linear regression makes the assumption that, at each level of the predictor (X), the distribution of the outcome Y (this is known as the conditional distribution of Y given X) is distributed Normal with ***means that lie on a line (simple linear regresssion) or a plane (multiple linear regression)***. By any means you like, produce graphs to assess the linearity of Y = glucose in age and the linearity of Y=glucose in BMI.

#3.
Y is distributed normal with constant variance. Another assumption of linear regression is that the distribution of the outcome Y at each level of the predictor is normal with constant variance.. When this assumption is met, ***the distribution of the residuals is distributed Normal with mean = 0 and constant variance***. Thus, assessment of this assumption involves examination of the residuals after fitting the model. Consider the model you fit in exercise #1. By any means you like, assess the assumption of normality of the residuals.

#4.
Y is distributed normal with constant variance Consider the model you fit in exercise #1. By any means you like test the null hypothesis of constancy of variance of the residuals.

#5.
Partial F-Test. In question 2, where you assessed normality of Y=glucose in the predictor X=BMI, the loess smoother suggested that, possibly, the relationship of Y=glucose to body mass index might be modeled better as a quadratic, namely with two predictors: BMI and BMI^2. Create a new predictor that is BMI^2. Then, perform a partial F test of the null hypothesis that, controlling for linearity in BMI, there is no additional statistical significance in BMI^2 in explaining the variability in outcomes.

#6.

Multicollinearity and the assessment of variance inflation. ***Multicollinearity*** is said to be present when the predictors are themselves linearly related. While some multicollinearity might be reasonably expected, if it is too extensive, each predictor on its own possesses too little independent information for the prediction of outcome. The result is regression coefficients with very large variances, or variance inflation. A measure of this is the variance inflation factor statistic, **VIF**. Briefly, to obtain the VIF for a particular predictor, that predictor is regressed on all the other predictors "i" and an R-squared is obtained. The VIF for the predictor is then obtained as follows. Values of VIF < 10 are considered acceptable (translation: no worries!):

$$VIF_i = \frac{1}{\sqrt{1 - R^2_{\text{regression of ith on all other predictors}}}}$$

By any means you like, produce a table of VIF values for the 3 predictors in the model you fit in exercise #1.

#7.

Partial regression plots (also called "added variables plot"). In a partial regression/added variable plot, the extra significance of a new predictor, controlling for the variables already in the model is examined by plotting the residuals of the new predictor on the control variables on the horizontal axis versus the residuals of Y on the control variables on the vertical axis. In this way, the influence of the control variables is "adjusted out". The slope of the scatter in this plot is a visual of the adjusted slope that will be obtained for the new predictor upon its inclusion in the model. Nice!

#8.

Model misspecification. ***Model misspecification*** can occur in a variety of ways; e.g., if some predictors are not modeled correctly (e.g., linearity in the predictor is insufficient) or important predictors are missing. The Ramsey test tests the null hypothesis the curent model is adequately specified. By any means you like, perform the Ramsey test for the model you fit in exercise #1.

#9.

Outliers. ***Outliers*** are observations that are unusual in the Y-sense. They may or may not influence the fitted model. But it's good to take a look. The Bonferroni test examines the largest studentized residual. For this particular studentized residual it peforms a t-test of the null hypothesis that it is not statistically significantly different from the other studentized residuals. By any means you like, assess the model you fit in exercise #1 with respect to outliers.

#10.

High leverage observations. ***High leverage observations*** are observations that are unusual in the X-sense. They may or may not influence the fitted model. By any means you like, assess model you fit in exercise #1 with respect to leverage.

#11.

Influential observations. ***Influential observations*** do impact the fit! Their inclusion in the model changes the estimated betas. There are several approaches to detect influential observations. Among the most commonly used is the calculation of ***Cook's distance***. Briefly, the Cook's distance is a summary measure of the discrepancy in the estimation betas in two models, one with the observation included and the other with the observation not included. A plot of study id versus Cook's distance makes their detection easy; simply look for spikes! Several thresholds/cutoffs have been suggested for the identification of influential observations. My suggested guidelines are these: (1) look at the plot first; where you see spikes, these observations may be influential (take care, however, to notice the range of Cook's distances by examining the y-axis scale provided); (2) A Cook's distance > 1 is worth exploring further; (2) A Cook's distance > .5 is of mild interest. By any means you like, construct a plot of Cook's distances for the model you fit in exercise #1.