Homework 3
Due Friday, July 15

1. Some least squares questions:

a) (Closely paraphrased from a question in the 1985 edition of SW) Occasionally, data are modeled with a regression equation with intercept = 0:

$$\mathbf{Y_i = \beta_1 x_i + \varepsilon_i,\ i=1,...,n}$$

Show that the least squares estimate of β_1 in this model is given by $\hat{\beta}_1 = \Sigma\ x_i y_i / \Sigma\ x_i^2$. Assuming standard linear regression assumptions with $\sigma^2 = \mathrm{Var}(\varepsilon_i)$, show that the estimator is unbiased and has variance = $\sigma^2 / \Sigma\ x_i^2$.

b) In the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, i=1,...,n, suppose that each x_i is replaced by:

i) $x_i + c$ (e.g, suppose that the covariate is measured with an instrument that has a nonrandom, constant calibration bias c)

ii) cx_i, $0<c<1$ (e.g, suppose that the covariate is systematically under-reported by a nonrandom, constant attenuation factor c)

Compare the least squares estimates of β_0 and β_1 for the original *x*s versus both of the replacement *x*s. For each replacement strategy and each of the estimates: please write down as precisely as you can what is the relationship between the estimate with the replacement *x*s and the estimate with the original *x*s. For example, the two may be equal; may differ by a constant that you specify; may be proportional but not equal, with one multiplying the other by a factor that you specify; etc. Please show your work.

c) In b) above, suppose that each x_i is replaced by $x_i + \delta_i$, where $\{\delta_i, i=1,...,n\}$ are mutually independent, identically-distributed, mean-0 RANDOM errors. Assume that the covariate error distribution does not depend on the *x*s and that the covariate errors are not correlated with the outcome errors (εs). How are the least squares estimates of β_0 and β_1 affected? You may approximate sample distribution characteristics by their population distribution counterparts, and justify as rigorously as possible.

d) In the multiple linear regression model $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$ suppose that each Y_i is replaced by a linear transformation $Y_i^* = a + cY_i$. For example, to translate our temperatures analysis from Fahrenheit to Celsius degrees, we would have a= -32*5/9 and c=5/9. How does the least squares estimate of $\underline{\beta}$ compare for outcomes $\underline{Y}$ versus $\underline{Y}^*$? How do confidence intervals for β_1 compare in the two models (say, one with Fahrenheit degrees and one with Celsius degrees)? Please show your work.

2. A data analysis: On the class website please find the dataset used to analyze Maximum January Temperatures in class. It has an added covariate: the altitude of each city (in feet: sorry!). Re-analyze the temperatures data

* converting Fahrenheit degrees to Celsius degrees: C = (F-32)*5/9,

* converting altitude feet to meters: m = feet*12/39.37

* including altitude as a covariate:

a) Generate a matrix scatterplot of temperatures, latitudes, longitudes, and altitudes. Describe the association of altitude with temperature, latitude and longitude.

b) Fit a simple linear regression (SLR) model of temperature on altitude and a multiple linear regression (MLR) model of temperature on latitude, longitude and altitude (a single term for each covariate). How does the temperature association with altitude compare in the two models? Identify the total, direct and indirect associations of temperature with altitude.

c) Is altitude associated with temperature independently of latitude and longitude? Provide an estimate with a measure of its uncertainty, and interpret it.

d) Examine plots of residuals versus fit and residuals versus altitude for both models (SLR, MLR). What do you observe? Do any of the assumptions appear to be violated?

e) Now use a linear spline with a single knot at altitude = 250 meters to model the relationship of temperature with altitude in the MLR. Does the relationship between mean temperature and altitude change at 250 meters? Provide an inference.

f) Create partial residual plot of temperature versus altitude for the model you fit in d). Describe the fitted relationship. Does it seem to well describe the data?

g) What is the fitted (predicted) temperature for Baltimore, Maryland based on your final model?

h) Report and interpret R^2 for your final model.

i) Summarize your results.