Theory Exam 2: Practice Exam

36-401, Modern Regression, Fall 2015

12 November 2015

This is longer than what we'd expect you to do in the exam, to give you more problems to practice on.

1. *Linear transformations preserve information* You have regressed Y on variables $X_1, X_2, \ldots X_p$. Your colleague, A. Y. K. Bob, has regressed Y on the variables $Z_1, Z_2, \ldots Z_p$, where

$$Z_j = c_{j0} + \sum_{k=1}^{p} c_{jk} X_k$$

That is, Bob has applied a linear transformation to the predictors (but not to the response).

   (a) Show that Bob's $n \times (p+1)$ design matrix $\mathbf{Z}$ is related to yours via

$$\mathbf{Z} = \mathbf{X}\mathbf{t}$$

   for some $(p+1) \times (p+1)$ matrix $\mathbf{t}$; explain how the entries in $\mathbf{t}$ are related to Bob's coefficients c.

   (b) Using the hat matrices of the two regressions, show that your fitted values and Bob's fitted values are exactly equal, if $\mathbf{t}$ is invertible.

   (c) Show that, $\widehat{\beta}$ is your vector of coefficients, and if $\mathbf{t}$ is invertible, then Bob's vector of coefficient estimates is exactly

$$\mathbf{t}^{-1}\widehat{\beta}$$

   (d) Is there any point to Bob's transformation of the predictor variables?

2. *Iterated regression* You have fitted a linear regression model by ordinary least squares. The diagnostic plots suggest that there might be some bias in the fitted values. Your boss suggests running a linear regression of your residuals on the predictors, and using its fitted values to correct the bias of the first regression. Show that the fitted values for this second regression will all automatically be zero. (*Hint:* Express the residuals of the first regression in terms of the hat matrix, and then use the fact that the hat matrix is idempotent.)

3. *Urban economies revisited* The data set on urban economies used for homework 3 contains, in addition to the per-capita gross metropolitan product (in dollars per person per year) and the population of each city, the fraction (not percentage) of each city's economy devoted to four industries: finance, "professional and technical" services, information and communications technologies (ICT), and management services. Here is the summary of a linear model fit to this data:

```
Call:
lm(formula = pcgmp ~ . - MSA, data = bea)

Residuals:
   Min     1Q Median     3Q    Max
-19160  -4813   -806   3087  25556

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.17e+04   1.88e+03   11.52  < 2e-16
pop         2.24e-03   1.09e-03    2.05  0.04249
finance     2.42e+04   1.18e+04    2.05  0.04252
prof.tech   3.09e+04   3.75e+04    0.82  0.41168
ict         6.42e+04   1.61e+04    3.98  0.00011
management  1.95e+05   8.58e+04    2.28  0.02440

Residual standard error: 7250 on 127 degrees of freedom
Multiple R-squared:  0.433,Adjusted R-squared:  0.411
F-statistic: 19.4 on 5 and 127 DF,  p-value: 2.54e-14
```

(a) Write the equation for the estimated model.

(b) What is the root mean squared error of the model?

(c) Provide an interpretation of the coefficient on `prof.tech`.

(d) Provide a 95% confidence interval for the coefficient on `prof.tech`, or explain what information you would need to calculate it that you are missing. (You may consult a table of the Gaussian, t, χ^2 or F distribution, as appropriate; on the exam, the relevant tables would be provided.)

(e) Based on the summary and/or your confidence interval, would it be reasonable to drop `prof.tech` from the model? (If you believe you cannot answer this without the confidence interval, and that you don't have the information to find the confidence interval, explain how you would use the CI to answer the question.)

(f) As of 2006 (when this data was collected), Pittsburgh[1] had the following values for all the variables:

[1] The whole metropolitan area, not just the legal city.

pop	finance	prof.tech	ict	management
2361000	0.2018	0.0777	0.03434	0.02946

What per-capita gross metropolitan product does the model predict for Pittsburgh?

(g) The leverage of Pittsburgh is 0.088. Does this give you enough information to calculate a standard error for the prediction for Pittsburgh? If yes, what is it? If not, what more do you need?

(h) Provide a 95% probability interval for Pittsburgh.

4. *Urban economies revisited continued* A simple regression of per-capita gross metropolitan product on population leads to the following summary:

```
Call:
lm(formula = pcgmp ~ pop, data = bea)

Residuals:
   Min     1Q Median     3Q    Max
-16339  -5557  -1570   4708  37566

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.94e+04   8.59e+02   34.19  < 2e-16
pop         5.86e-03   9.98e-04    5.87  3.3e-08

Residual standard error: 8440 on 131 degrees of freedom
Multiple R-squared:  0.208,Adjusted R-squared:  0.202
F-statistic: 34.5 on 1 and 131 DF,  p-value: 3.32e-08
```

(a) `pop` has a larger coefficient in the simple regression than in the larger model. Give at least one possible explanation of this. *Hint:* Think about the signs of the coefficients.

(b) What is the difference in MSEs between the two models?

(c) Carefully state the null hypothesis for a partial F test comparing this model to the model in the previous problem.

(d) What is the F statistic for a partial F test of the joint significance of all the industry-share coefficients? What is the p-value?

(e) What is the difference in the Mallow's C_p statistic for the two models? Which model is preferred by that criterion?