Statistics GU4205 — Fall 2016
Final Exam

December 19, 2016

NAME: UNI:

Instructions: Write your name and UNI in the spaces provided above. Do not begin work until instructed to do so.

You have 170 minutes to complete this examination. Read each part of each question carefully. There are a total of 80 points on this exam — you are responsible for checking that your paper is complete. You are permitted two $8\frac{1}{2}$ by 11 sheets (four sides total) of handwritten notes and a hand-held calculator. **No other outside material or assistance is permitted.**

Section 1: True or false? Circle the appropriate choice (1 point each).

1. Consider a joint distribution in which $\mathrm{E}[\log_2 Y | X = x] = \beta_0 + \beta_1 \log_2 x$. Under this model, doubling the value of x produces a $4\beta_1$ multiplicative effect on the expected value of Y given $X = x$.

   TRUE FALSE

2. If we take the residuals from the regression of y on x_1, and plot them versus the residuals from regressing x_2 on x_1, the slope for the least squares regression line on this plot will be exactly equal to b_2, the estimated coefficient of x_2 in the multiple regression of y on x_1 and x_2.

   TRUE FALSE

3. If we take the residuals from the regression of y on x_1, and plot them versus the residuals from regressing x_2 on x_1, the intercept for the least squares regression line on this plot will be exactly equal to zero.

   TRUE FALSE

4. The coefficient of multiple determination, denoted R^2, gives the proportionate reduction to the total variation in the response variable Y that is achieved by accounting for the linear association between Y and the predictor variables in a regression.

True False

5. Let $\max_j r_j^2$ and $\min_j r_j^2$ denote the maximum and minimum values of the squared correlation coefficients between the response variable Y and the predictor variables $X_1, \ldots, X_{p-1}$; then the coefficient of multiple determination R^2 satisfies $\min_j r_j^2 \leq R^2 \leq \max_j r_j^2$.

True False

6. In a multiple linear regression with $p-1$ predictor variables and an intercept term, and assuming a constant variance, an unbiased estimator of that variance is the mean squared error, $MSE = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

True False

7. Changing the order in which terms are entered into a multiple linear regression model will change the decomposition of the regression sum of squares, but not their sum. With two predictors, for example, although in general $SSR(X_1) \neq SSR(X_1|X_2)$, and $SSR(X_2|X_1) \neq SSR(X_2)$, it will still be the case that

$$SSR(X_1) + SSR(X_2|X_1) = SSR(X_2) + SSR(X_1|X_2) = SSR(X_1, X_2) \ .$$

True False

8. Suppose we regress Y on X_1, save the residuals as $e(Y|X_1)$; and likewise regress X_2 on X_1, save those residuals as $e(X_2|X_1)$. Then the squared correlation between these two sets of residuals will equal

$$R^2_{Y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)} \ ,$$

the *coefficient of partial determination* for X_2, accounting for X_1.

True False

9. The coefficient of partial determination $R^2_{Y2|1}$ will not in general be equal to the corresponding simple coefficient of determination R^2_{Y2}, but the decomposition of the coefficient of multiple determination holds that $R^2_{Y1} + R^2_{Y2|1} = R^2_{Y2} + R^2_{Y1|2}$.

True False

10. Adding a quadratic term to a simple linear regression model may be appropriate when the optimal response value is attained at an intermediate value of the predictor, that is, when the dose-response relationship is not monotone. The additional term in the mean function may have the effect of reducing R^2, but that sacrifice is justified by the improved realism of the model.

    TRUE FALSE

Note: Questions 11–15 are worth 2 points each.

If you answer false, you must indicate what number would make the statement true.

11. Consider a regression problem with two quantitative predictors. The mean function for the general second-order model (quadratic in both terms, plus interaction modeled as the product x_1x_2) will have five parameters.

    TRUE FALSE

12. Consider a regression problem with two predictor variables, both categorical: one has four levels, and the other has five levels. The mean function for the *main effects* model (no interaction) will have eight parameters.

    TRUE FALSE

13. Consider a regression problem with two predictor variables, both categorical: one has four levels, and the other has five levels. The mean function for the second order model (main effects and interaction) will have 21 parameters.

    TRUE FALSE

14. Consider a regression problem with two predictor variables, one is categorical with four levels, and the other is continuous. The mean function for the *parallel linear regressions* (first order) model will have five parameters.

    TRUE FALSE

15. Consider a regression problem with two predictor variables, one is categorical with four levels, and the other is continuous. The mean function for the *separate linear regressions* (first order plus interaction) model will have 10 parameters.

    TRUE FALSE

16. Consider a linear regression on two predictor variables x_1 and x_2. Other things being equal, the greater the difference in their average values, i.e., the greater is $|\bar{x}_1 - \bar{x}_2|$, the greater is the variance of the least squares estimates of both their coefficients.

TRUE FALSE

17. Consider a linear regression on two predictor variables. Other things being equal, the more highly correlated the predictors are to each other, the greater is the variance of the least squares estimates of both of their coefficients.

TRUE FALSE

18. Consider a linear regression on two predictor variables. If the predictor variables are *perfectly* correlated (collinear), then the least squares estimation of their coefficients won't even have a unique solution.

TRUE FALSE

19. Consider a linear regression on $p-1$ predictors, plus intercept. The variance inflation factor for the jth predictor x_j is a function of the coefficient of multiple determination from regressing x_j on $x_1, \ldots, x_{j-1}$; thus the variance inflation factors for a set of predictors will depends on what order the terms are listed in the model specification.

TRUE FALSE

20. If the Schwartz Bayesian Criterion, or *BIC*, selects a different set of predictors than does *AIC*, it will generally be the case that the model favored by *BIC* is simpler (contains fewer terms); this follows since *AIC* contains a more severe penalty for model complexity.

TRUE FALSE

21. If data are available for $P-1$ predictor variables, there are 2^{P-1} potential first-order models containing an intercept term. For five predictors ($2^5 = 32$) or even ten ($2^{10} = 1024$) it may not be unreasonable for a computer routine to search them all. Much more than that, however, and some sort of search strategy (e.g., stepwise regression) will be necessary (or at least desirable).

TRUE FALSE

22. Depending on the number of terms in the optimal subset of predictor variables, one of *backward selection* and *forward elimination* will be more efficient than the other. But either algorithm will identify and select this "best model" eventually.

True False

23. A simplistic but useful approach to testing for lack of fit in multiple linear regression is to consider the model

$$\mathrm{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1}x_{p-1} + \beta_{jj}x_j^2$$

and test $H_0 : \beta_{jj} = 0$ versus $H_A : \beta_{jj} \neq 0$. Do this for each $j = 1, \ldots, p-1$.

True False

24. If we conduct the *tests for curvature* described above, as well as *Tukey's test for nonadditivity*, and fail to reject H_0 in every single instance, then we can safely conclude that the first order (linear) mean function is correct.

True False

25. In the *Breusch-Pagan test for nonconstant variance*, the alternative hypothesis holds that the variance function depends on a linear combination of the predictor variables, specifically that

$$\mathrm{Var}[Y|\boldsymbol{X} = \boldsymbol{x}] = \exp\{\gamma_0 + \gamma_1 x_1 + \cdots + \gamma_{p-1}x_{p-1}\} \ .$$

The null hypothesis holds that $\gamma_1 = \cdots = \gamma_{p-1} = 0$.

True False

26. Suppose we reject H_0 in a Breusch-Pagan test, and conclude that the assumption of constant variance does not hold for our data. Some possible courses of action (not an exhaustive list) would be (1) investigate the need to transform the response variable; (2) consider weighted least squares; and (3) do nothing: ordinary least squares estimation of the mean function parameters will still be unbiased (if no longer optimal), and inferences will still be approximately valid.

True False

27. In a regression problem with response variable y, and $\boldsymbol{x}$ a vector of predictors, an *outlier* is characterized by an unusual y-value, given its $\boldsymbol{x}$-value; it is quite possible that y_j sits at or near the center of the marginal distribution of the y_i's, but case j could still be an outlier (depending on the value of $\boldsymbol{x}_j$).

TRUE FALSE

28. In a regression problem with response variable y, and $\boldsymbol{x}$ a vector of predictors, a high-leverage case is characterized by an unusual $\boldsymbol{x}$-value. In fact, the formula for computing the "hat value" h_{jj} (the usual measure of case j's leverage) does not depend on y_j at all.

TRUE FALSE

29. One approach to assessing the "outlierness" of case j in a linear regression problem involves fitting the model

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \delta_j I_{\{i=j\}} + \varepsilon_i$$

for $i = 1, \ldots, n$. Since the estimation of $\boldsymbol{\beta}$ is unchanged by the additional term in the above model, the estimate of δ_j is just the case j residual, $e_j = y_j - \mathbf{x}_j'\mathbf{b}$.

TRUE FALSE

30. When we conduct an outlier t-test for only the most "outlier-ish" case in a data set, we have implicitly conducted n such tests, and selectively reported only the most striking result. A valid (in fact conservative) adjustment for this *multiple testing* issue is the so-called *Bonferroni correction*, in which we simply multiply the p-value by n (capping at 1 of course).

TRUE FALSE

31. It is mathematically possible (though rarely encountered in practice) for a case whose residual and hat-values are both close to zero to still be an *influential case*, in the sense of having the highest (or near highest) *Cook's distance* in a data set.

TRUE FALSE

32. Under the statistical model $Y_i \sim$ indep Normal $\left(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2\right)$ for $i = 1, \ldots, n$, the maximum likelihood estimator of $\boldsymbol{\beta}$ is the value which minimizes

$$Q\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 \ ;$$

that is, under normality, maximum likelihood and least squares are equivalent criteria.

True False

33. Under the statistical model $Y_i \sim$ indep Normal $(\mathbf{x}_i'\boldsymbol{\beta}, k/w_i)$ for $i = 1, \ldots, n$, where the w_i are known and not all equal, the maximum likelihood estimator of β is the value which minimizes

$$Q\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 \ ;$$

that is, under normality, even with unequal variance, maximum likelihood and ordinary least squares are equivalent criteria.

True False

34. One of the strengths of the weighted least squares approach to regression estimation is that it requires no assumption about the relative variance for different observations of the response variable.

True False

35. The basic idea of the weighted least squares approach can be summarized as: Those observations which are subject to greater variability should be given greater weight in the model estimation. (And those observations subject to less variability should be given lower weight.)

True False

End of Section 1.

Section 2: Answer all questions in the space provided, use additional pages (bluebook) if necessary.

1. (20 points) A national insurance organization wants to study the consumption pattern of cigarettes in all 50 states and the District of Columbia. They fit a linear regression of *Sales* (packs of cigarettes sold per capita in each state) on average *Price* (in cents) of a pack of cigarettes, per capita *Income*, percentage of adults to complete *HS*, and median *Age* in each state. Estimated regression coefficients and a sequential ANOVA table are given below.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  88.55616   69.95168   1.266  0.21190
Price        -3.40795    0.99371  -3.430  0.00129
Income        0.02300    0.00847   2.715  0.00929
HS           -0.46035    0.60566  -0.760  0.45108
Age           3.66425    2.30142   1.592  0.11820
```

```
           Df  Sum Sq  Mean Sq  F value    Pr(>F)
Price       1    4648   4647.5   6.0414  0.017801
Income      1    8223   8223.2  10.6895  0.002044
HS          1    1218   1218.0   1.5833  0.214629
Age         1    1950   1950.1   2.5350  0.118196
Residuals  46   35387    769.3
```

(a) Identify and interpret the estimated regression coefficient for the *Price* term: What does this number mean? Can we conclude that a sufficiently large cigarette tax (increasing the price of a pack of cigarettes), will reduce a state's cigarette sales? Briefly explain.

(b) Assess whether the term *HS* can reasonably be dropped from the model by reporting the P-value for a test of the null hypothesis that its coefficient is zero. Your conclusion?

(c) Assess whether *HS* and *Age* can both be dropped from the model, by testing the null hypothesis that their coefficients are simultaneously zero. You should: (i) compute a test statistic, and (ii) specify its null distribution, i.e., the distribution to which this value is compared in order to obtain a P-value. (You do not have to actually obtain that P-value.)

(d) Of the variation in *Sales* that remains after accounting for its linear association to *Price* and *Income*, what proportion remains unexplained after accounting for *HS* and *Age* as well?

2. (10 points) Researchers wishing to study teenage gambling in Great Britain took a survey of $n = 47$ British teenagers, then fit the regression of

$$y = \ln(\text{expenditure on gambling in pounds per year})$$

on $x_1 =$ socioeconomic *status* score (based on parents' occupation), $x_2 = \ln(\textit{income})$ (in pounds per week), $x_3 = \textit{verbal}$ score in words out of 12 correctly defined, and $x_4 = \textit{sex}$ (0=male, 1=female). The estimated regression coefficients were:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.46076    1.62374   0.900  0.37345
status       0.04509    0.02579   1.748  0.08774
ln.income    1.26811    0.45022   2.817  0.00736
verbal      -0.48559    0.20386  -2.382  0.02182
sex         -1.78504    0.74839  -2.385  0.02166
```

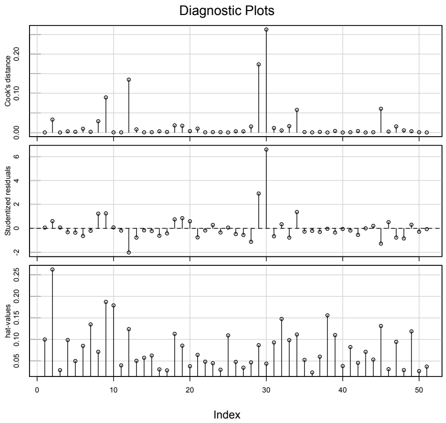
(a) Consider two British teenagers: a girl with *status* $= 30$ and *verbal* $= 4$, and a boy with *status* $= 60$ and *verbal* $= 6$ (assume their respective incomes are equal). Use the model fit above to make a prediction of which child gambles more and by how much. Provide as precise an answer as you can, for a client who does not understand logarithms.

(b) Can you compute a standard error for your estimate in part (a)? A portion of the estimated variance-covariance matrix $\widehat{\text{Cov}}(\mathbf{b})$, from the R function `vcov()`, is given below.

```
        status  verbal     sex
status  0.0007 -0.0028  0.0101
verbal -0.0028  0.0416 -0.0304
sex     0.0101 -0.0304  0.5601
```

3. (10 points) A national insurance organization wanted to study the consumption pattern of cigarettes in all 50 states and the District of Columbia. They fit a linear regression of $y =$ packs of cigarettes sold per capita in each state on $x_1 =$ median age, $x_2 =$ per capita income, and $x_3 =$ average price of a pack of cigarettes.

   Index plots for Cook's distance, outlier t-statistics, and hat values are given below.



   (a) With a single sentence, explain to someone who has zero knowledge of or interest in statistical modeling: What is it about Nevada and New Hampshire (cases 29 and 30, respectively) that makes them stick out in these plots?

   (b) Carefully explain how you could generalize the idea of *Cook's distance* to compute a measure of the combined influence of cases 29 and 30. Can you foresee a practical difficulty in the assessment of this measure?