Practice Problems for Final

401 GSI team

4/12/2018 and 4/13/2018

Sanity Checking Data

Car Drivers

Car drivers adjust the seat position for comfort. We are interested in determining the major factors that determine the seat position (measured as hip center in mm) and have collected the following information on 38 drivers: age in years, weight in lbs, and height with shoes, height without shoes, seated height, lower arm length, thigh length, lower leg length all in cm.

```r
library(faraway)
library(ggplot2)
data("seatpos")
head(seatpos)
```

```
##   Age Weight HtShoes    Ht Seated  Arm Thigh  Leg hipcenter
## 1  46    180   187.2 184.9   95.2 36.1  45.3 41.3  -206.300
## 2  31    175   167.5 165.5   83.8 32.9  36.5 35.9  -178.210
## 3  23    100   153.6 152.2   82.9 26.0  36.6 31.0   -71.673
## 4  19    185   190.3 187.4   97.3 37.4  44.1 41.0  -257.720
## 5  23    159   178.0 174.1   93.9 29.5  40.1 36.9  -173.230
## 6  47    170   178.7 177.0   92.4 36.0  43.2 37.4  -185.150
```

To determine these factors, we start by fitting the following linear model:

```r
comfort <- lm(hipcenter ~ ., data = seatpos)
summary(comfort)
```

```
##
## Call:
## lm(formula = hipcenter ~ ., data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.827 -22.833  -3.678  25.017  62.337
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 436.43213  166.57162   2.620   0.0138 *
## Age           0.77572    0.57033   1.360   0.1843
## Weight        0.02631    0.33097   0.080   0.9372
## HtShoes      -2.69241    9.75304  -0.276   0.7845
## Ht            0.60134   10.12987   0.059   0.9531
## Seated        0.53375    3.76189   0.142   0.8882
## Arm          -1.32807    3.90020  -0.341   0.7359
## Thigh        -1.14312    2.66002  -0.430   0.6706
## Leg          -6.43905    4.71386  -1.366   0.1824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.72 on 29 degrees of freedom
## Multiple R-squared:  0.6866, Adjusted R-squared:  0.6001
## F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

1. Explain why the intercept is the only significant variable but overall the model is highly significant.

2. Suppose we instead fit the following model, $y_i = b_0 + b_1 x_{i1}$, where y_i is the hipcenter of person i and x_{i1} is the height of person i. How would you expect the standard errors to differ between this model and the one fit above?

Garbage Data

Suppose we are interested in determining the energy content of municipal waste. This information is useful for the design and operation of municipal waste incinerators. The variables are `Energy` - energy content (kcal/kg), `Plastic` - % plastic composition by weight, `Paper` - % paper composition by weight, `Garbage` - % garbage composition by weight, and `Water` - % moisture by weight.

A number of samples of municipal waste were obtained, and a regression to predict the `Energy` was computed and the following output obtained:

```
Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)   2245.09      177.89    12.62    2.4e-12
Plastics        28.92        2.82    10.24    2.0e-10
Paper            7.64        2.31     3.30     0.0029
Garbage          4.30        1.92     2.24     0.0340
Water          -37.36        1.83   -20.37     <2e-16

Residual standard error: 31.5 on 25 degrees of freedom
Multiple R-Squared: 0.964, Adjusted R-squared: 0.958
F-statistic: 168 on 4 and 25 DF, p-value: <2e-16
```

1. Suppose sample A has 10% more plastic (in absolute terms) than sample B. What is the predicted energy content of sample A compared to sample B? Give the mathematical formula for calculating a 95% prediction interval for this difference. Substitute all possible values from the regression output.

2. What energy content would this model predict for a sample that was purely water? Comment on the reliability of this prediction. What is the term for a prediction like this.

3. Compute the 95% confidence interval for the coefficient of garbage. Substitute all possible values from the regression output.

4. How is the residual standard error calculated for this model? (Give a formula.)

[Source: This question is adapted from *Applied Statistics for Engineers and Scientists* by Jay L. Devore, Nicholas R. Farnum, Jimmy A. Doi (Ch. 3, Question 56).]

Education Study Data

Recall the National Education Longitudinal Study of 1988 which examined schoolchildren's performance on a math test score in 8th grade. "ses" is the socioeconomic status of parents and "paredu" is the parents highest level of education achieved (less than high school, high school, college, BA, MA, PhD).

```
library(MASS)
data(nels88)

ses_edu_lm <- lm(math ~ ses + paredu + paredu:ses, data = nels88)
```

1. Name 2 model diagnostics would you perform to check if this is a valid model.
2. Name one potential confounding variable of this dataset.
3. Suppose we fit the following model: $y_i = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$ where y_i is the math score, x_1 is the socioeconomic status of parents, and x_2 is the parents' highest level of education.

a) Does it initially make sense to include the interaction terms? Explain.

b) We want to know if the interaction terms are significant. Interpret the ANOVA below to answer this question. Provide the null and alternative models in subscript form, the test statistic, and the distribution of the test statistic under the null hypothesis in explaining your answer. You may define the residual sum of squares in words.

```
anova(ses_edu_lm)
```

```
## Analysis of Variance Table
##
## Response: math
##             Df  Sum Sq Mean Sq  F value    Pr(>F)    
## ses          1 12391.4 12391.4 173.5021 < 2.2e-16 ***
## paredu       5  1642.4   328.5   4.5994 0.0004959 ***
## ses:paredu   5   370.7    74.1   1.0381 0.3957126    
## Residuals  248 17712.0    71.4                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Theoretical Exercises

1. a) Suppose we have the following matrix $\mathbb{X} = \begin{bmatrix} 36 & 24 & 12 \\ 24 & 24 & 0 \\ 12 & 0 & 24 \end{bmatrix}$. Find a vector α s.t. $\mathbb{X}\alpha = \mathbf{0}$.

   b) Suppose we have the following matrix $\mathbb{X} = \begin{bmatrix} 1 & 5 & 2 \\ 1 & 7 & 3 \\ 1 & 15 & 7 \end{bmatrix}$. Find a non-zero vector α s.t. $\mathbb{X}\alpha = \mathbf{0}$.

   c) What can you say about the collinearity of these two matrices?

2. Explain why we might be able to infer causation as opposed to just correlation in a regression of a designed experiment.

For additional exercises, see notes, quizzes, and homeworks.