Quiz 2

Modern Data Mining/Linda

March 1, 2022

Name: _____
Section (471, 571, 701): _____

This is an open book, 10-minute quiz. Choose the correct answer(s). There might be more than one right answer in some questions. No calculations are needed.

Insurance companies try to charge a higher premium than the amount paid to the insured. For this reason, insurance companies invest a lot of time, effort, and money in creating models that are able to accurately predict health care costs. In order to fulfill this mission, we would build an adequate model and optimize its performance.

We have a dataset that includes 1338 observations on 7 variables:

- `age`: The age of the policy holder.
- `sex`: The sex of the policy holder. Values: (`male`, `female`)
- `bmi`: The Body Mass Index of the policy holder. BMI gives an understanding of body weights that are relatively high or low relative to height.
- `children`: The number of dependents the policy holder has.
- `smoker`: Whether the policy holder smokes. Values: (`yes`, `no`)
- `region`: Part of the US the policy holder lives in. Values: (`northeast`, `southeast`, `southwest`, `northwest`)
- `charges`: Medical costs billed to the policy holder.

Our goal of the study is to understand how each variable affects the response `charges` and build a good linear model to predict `charges`. We will only include the non-smokers and convert the variables `sex` and `region` to categorical variables; the rest of variables is regarded as continuous type. We use `lcharges = log(charges)` as the response.

```
url <- 'https://raw.githubusercontent.com/stedy/Machine-Learning-with-R-datasets/master/insuranc
df_all <- read_csv(url, col_types = cols(.default="d", sex="f", smoker="f", region="f"))
df <- df_all %>%
  filter(smoker == "no") %>%
  select(-smoker) %>%
  mutate(lcharges = log(charges))
head(df)
```

```
## # A tibble: 6 x 7
##     age sex      bmi children region    charges lcharges
##   <dbl> <fct>  <dbl>    <dbl> <fct>       <dbl>    <dbl>
## 1    18 male    33.8        1 southeast   1726.     7.45
## 2    28 male    33          3 southeast   4449.     8.40
## 3    33 male    22.7        0 northwest  21984.    10.0
## 4    32 male    28.9        0 northwest   3867.     8.26
## 5    31 female  25.7        0 southeast   3757.     8.23
## 6    46 female  33.4        1 southeast   8241.     9.02
```

Simple Regression

We first run a linear regression of `lcharges` on `region` using `df`. Recall that there are 4 regions: `northeast`, `southeast`, `southwest`, `northwest`.

```
fit1 <- lm(lcharges ~ region, data = df)
summary(fit1)
```

```
##
## Call:
## lm(formula = lcharges ~ region, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.674 -0.497  0.118  0.550  1.810
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.7525     0.0454  192.84   <2e-16 ***
## regionsoutheast  -0.0555     0.0638   -0.87    0.385
## regionnorthwest   0.0618     0.0642    0.96    0.336
## regionnortheast   0.1428     0.0648    2.20    0.028 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.74 on 1060 degrees of freedom
## Multiple R-squared:  0.00976,    Adjusted R-squared:  0.00695
## F-statistic: 3.48 on 3 and 1060 DF,  p-value: 0.0155
```

1. True or False? Based on the summary table of `fit1`, we CANNOT reject the null hypothesis that the average `lcharges` for policy holder from southwest region is the same as that from northeast origin under significance level $\alpha = .05$.

(A) TRUE

(B) FALSE

Answer: (A). `southwest` **is the base level in this analysis. The coef for** `northeast` **is the mean**

difference of `lcharges` **between** `northeast` **and** `southwest`**.**

2. According to the `fit1`, what is the average `lcharges` for policy holders from southeast?

(A) 8.75

(B) -0.06

(C) 8.75 -0.06

The answer is (C).

We then fit a linear regression model of `lcharges` on `age`.

```
fit2 <- lm(lcharges ~ age, data = df)
summary(fit2)
```

```
##
## Call:
## lm(formula = lcharges ~ age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8671 -0.1940 -0.0765  0.0497  2.2562
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.133743   0.041150   173.4   <2e-16 ***
## age         0.042008   0.000984    42.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.45 on 1062 degrees of freedom
## Multiple R-squared:  0.632,  Adjusted R-squared:  0.632
## F-statistic: 1.82e+03 on 1 and 1062 DF,  p-value: <2e-16
```

3. Based on summary of `fit2`, choose correct answer(s)

(A) Age is statistically significant at level $\alpha = .05$.

(B) On average, a 25-year-old male is likely to be charged 0.04 more than a 24-year-old female.

The answer is both (A) and (B) since `age` **is not included in the model.**

Building a better model

We construct a linear model `lcharges` including more covariates.

```
fit3 <- lm(lcharges ~ age + sex + children + region, data=df)
summary(fit3)
```

```
##
## Call:
## lm(formula = lcharges ~ age + sex + children + region, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6261 -0.1631 -0.0712 -0.0162  2.3735
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.985183   0.047443  147.23  < 2e-16 ***
## age              0.041568   0.000907   45.83  < 2e-16 ***
## sexmale         -0.101575   0.025518   -3.98  7.3e-05 ***
## children         0.128524   0.010481   12.26  < 2e-16 ***
## regionsoutheast  0.017027   0.035827    0.48    0.635
## regionnorthwest  0.107260   0.036008    2.98    0.003 **
## regionnortheast  0.181534   0.036356    4.99  6.9e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.42 on 1057 degrees of freedom
## Multiple R-squared:  0.69,   Adjusted R-squared:  0.688
## F-statistic:  391 on 6 and 1057 DF,  p-value: <2e-16
```

4. True or False? Based on the summary of `fit3`, because `regionsoutheast` and `regionnorthwest` both have p-values larger than 0.001, we conclude at level $\alpha = .001$ that there is no effect of policy holder's region over insurance premium in this model.

(A) TRUE

(B) FALSE

**The answer is (B).

5. True or False? Based on the summary of `fit3`, even though region `southeast` is not significant at level $\alpha = .05$, it CAN be statistically significant when we account for more variables.

(A) TRUE

(B) FALSE

The answer is (A).

6. Choose the correct answer. Based on `fit3`, we would like to estimate the mean of `lcharges`

for a customer with the following measurement: `age = 50, sex = Female, bmi = 30, children = 5` and `region = southwest`.

(A) We can not do it since `bmi` is not included in `fit3`.

(B)
$$6.99 + 50 \times 0.04 + 5 \times 0.13$$

(C)
$$6.99 + 50 \times 0.04 + 5 \times 0.13 + 0.02$$

(D)
$$6.99 + 50 \times 0.04 - 0.1 + 5 \times 0.13$$

The answer is (C).

We extend the previous model to contain additional interaction terms.

```
fit4 <- lm(lcharges ~ age + sex + bmi + children  + region +
             age*sex + bmi*sex, data=df)
Anova(fit4)
```

```
## Anova Table (Type II tests)
##
## Response: lcharges
##           Sum Sq   Df F value  Pr(>F)
## age          354    1 2065.84 < 2e-16 ***
## sex            3    1   16.09 6.5e-05 ***
## bmi            0    1    0.30  0.5836
## children      27    1  154.99 < 2e-16 ***
## region         6    3   11.06 3.7e-07 ***
## age:sex        2    1   10.64  0.0011 **
## sex:bmi        0    1    0.17  0.6765
## Residuals    181 1054
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7. True or False? Based on the `Anova` table of `fit4`, `region` is significant at 0.05 level.

(A) TRUE

(B) FALSE

The answer is (A).

8. We want to estimate the prediction error of model `fit4`. Choose the correct metric(s).

(A) Mallow C_p

(B) BIC

(C) Both (A) and (B).

(D) Neither (A) nor (B).

The answer is (A). Once again C_p estimates testing errors.

9. We implement the best subsets regression. (we use a different package to calculate the Cp but it does the same thing as regsubsets).

```
selection_result <- olsrr::ols_step_best_subset(fit4)

as_tibble(selection_result) %>%
  transmute(model=mindex, predictors, Mallow_Cp=cp)
```

```
## # A tibble: 7 x 3
##   model predictors                                    Mallow_Cp
##   <int> <chr>                                             <dbl>
## 1     1 age:sex                                          196.
## 2     2 children age:sex                                  49.3
## 3     3 children region age:sex                           18.8
## 4     4 sex children region age:sex                        0.475
## 5     5 sex children region age:sex sex:bmi                2.00
## 6     6 sex bmi children region age:sex sex:bmi            4.00
## 7     7 age sex bmi children region age:sex sex:bmi        6
```

Based on Mallow's C_p, which model has the smallest prediction error?

(A) model 2

(B) model 4

(C) model 6

(D) model 8

The answer is (B).