HW 4

Evan Cranmer and Xiaoqing Liu

2024-09-25

Examining medv

```
library(MASS)
data("Boston")
?Boston
#Examine medv as a function of crim, zn and indus in a multiple linear regression
medvmodel <- lm(medv ~ crim + zn + indus, data = Boston)

summary_medv <- summary(medvmodel)
```

1a and 1b looking at statistically significant predictors

To examine the significance of each predictor, we're going to compare our p-values from the medv model to the alpha value of 0.05. Then R will print out our decision based on whether our p-value is less than our alpha value.

```
#storing p-values of each predictor
p_values <- summary_medv$coefficients[, "Pr(>|t|)"]
p_values
```

```
##   (Intercept)          crim            zn         indus
## 8.475305e-122  2.515185e-08  8.889412e-04  1.765577e-10
```

```
alpha <- 0.05 # our alpha value

#printing whether we reject our null of fail to reject our null
rejection <- ifelse(p_values < alpha, "Reject H0", "Fail to Reject H0")

#storing our results in a data frame
significance_table <- data.frame(
  Predictor = rownames(summary_medv$coefficients),
  p_value = p_values,
  Decision = rejection
)

significance_table
```

```
##                Predictor       p_value  Decision
```

```
## (Intercept) (Intercept) 8.475305e-122 Reject H0
## crim               crim  2.515185e-08 Reject H0
## zn                   zn  8.889412e-04 Reject H0
## indus             indus  1.765577e-10 Reject H0
```

1b null and alternative hypotheses

Our null hypothesis states that our p-value is not statistically significant, H_0: p > alpha at 0.05. Our alternative hypothesis states that our p-value is statistically significant, H_0: p < alpha 0.05.

We found that all predictors are statistically significant, because their p-values are less than the alpha level of 0.05. We can reject the null hypothesis.

1c interpreting the regression coefficients

Crim: For every one unit increase in per capita crime rate, the expected value of median home value decreases by 0.24863 in the $1000s, holding other variables constant.

zn: For every one unit increase in proportion of residential land zoned for lots over 25,000 sq.ft., the expected value of median home value increases by 0.05850 in the $1000s, holding other variables constant.

indus: For every one unit increase in proportion of non-retail business acres per town., the expected value of median home value decreases by .41558 in the $1000s, holding other variables constant.

1d

Interpreting all predictors as the exposure of interest could be problematic because control variables could have interactions with other variables, or just simply not have the equal effects.

1E

```r
confint(medvmodel, level = 0.95)
```

```
##                  2.5 %      97.5 %
## (Intercept) 25.6954863 29.09380729
## crim        -0.3348958 -0.16236084
## zn           0.0241252  0.09287644
## indus       -0.5408945 -0.29026116
```

The 95% confidence intervals for the predictors are the following:

crim: (-0.334895, -0.16236084). If we were to take repeated samples of the population and calculate coefficient crim, the true value would be contained in the confidence interval (-0.334895, -0.16236084) 95% of the time.

zn: (0.0241252, 0.09287644). If we were to take repeated samples of the population and calculate coefficient zn, the true value would be contained in the confidence interval (0.0241252, 0.09287644) 95% of the time.

indus: (-0.5408945, -0.29026116). If we were to take repeated samples of the population and calculate coefficient indus, the true value would be contained in the confidence interval (-0.5408945, -0.29026116) 95% of the time.

Since the three variable confidence intervals do not pass through 0, all three are statistically significant and have a non-zero effect. So we would reject the null hypothesis in 1b. In 1a+1b we found that the p-values were less than .05, meaning the variables were statistically significant, and we can reject the null hypothesis that the variables are equal to zero. This corresponded to a confidence interval that did not include zero, which we just found.

1F

$R^2 = \frac{SS_{Reg}}{SS_Y} = 1 - \frac{RSS}{SS_Y}$ and $R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-p}$

```r
#storing the observed values for medv
y_medv <- Boston$medv

#storing predicted values from our medv model
y_medv_hat <- predict(medvmodel)

#calculating RSS
rss <- sum((y_medv - y_medv_hat)^2)

#finding SSy
ssy <- sum((y_medv - mean(y_medv))^2)

medv_r2 <- 1 - (rss/ssy)
print(medv_r2)
```

```
## [1] 0.2937136
```

```r
########### Now finding R2 adjusted

#storing number of observations
n <- length(y_medv)

#number of predictors not including the intercept
p <- 3

#finding our
medv_r2_adj <- 1 - ((1- medv_r2)*(n-1) / (n-p-1))
print(medv_r2_adj)
```

```
## [1] 0.2894927
```

Interpreting R squared and Adjusted R squared results

Our R-squared value for this model is approximately .294. Which means that approximately 29.4% of the variance in the dependent variable is explained by the independent variables.

Our adjusted R-squared value for this model is approximately .289. Which means that approximately 28.9% of the variance in the dependent variable is explained by the independent variables, and the value was adjusted for the number of parameters in our model.

2 Fitting a new model

```r
#creating new model with zn as only predictor
medvmodel2 <- lm(medv ~ zn, data = Boston)

#performing global f test
anova(medvmodel2, medvmodel)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ zn
## Model 2: medv ~ crim + zn + indus
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    504 37167
## 2    502 30170  2    6996.6 58.209 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After performing a global f test comparing both of the models, we find that the p-value for testing $H_0 : RSS_{Reduced} = RSS_{Full}$, is $< .001$. We can reject H_0 and conclude that the full model is better, because it has a significantly smaller RSS.