Stats 110 HW2

Viraj Vijaywargiya

2022-10-17

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

1)1a) False. **1b)** False. **1c)** True. **1d)** False. **1e)** True. **1f)** False.

2)2a) Same, as correlation between X and Y is the same as the correlation between Y and X. **2b)** Same, as R is the same, R-squared will be the same. **2c)** Different, as B1 is the slope and as Y and X change, B1 will differ. **2d)** Different, same reason as part c. **2e)** Same, as test statistic depends only on R and since R is the same, the test statistic will be the same. **2f)** Same, as residuals for each observation stays the same as the relation between X and Y stays the same.

3)

```
# Import MidWestSales dataset
    MidWestSales = read.table("/Users/virajvijaywargiya/Downloads/MidwestSales.txt", fill=TRUE, header=
    # This dataset dsigma(e)s not have names, so we will add names to the variables
    names(MidWestSales)=c("id","price","sqft","bed","bath","ac","garage","pool","year","quality","style
    view(MidWestSales)
```

3a)

```
lm(formula = price ~ sqft, data = MidWestSales)
```

```
##
## Call:
## lm(formula = price ~ sqft, data = MidWestSales)
##
## Coefficients:
## (Intercept)         sqft
##      -81433          159
```

Estimated regression equation: Y = -81433 + 159X.

3b) As the square footage of the house increases by 1 unit, the expected sale price increases by 159 dollars.

3c)

```
model = lm(formula = price ~ sqft, data = MidWestSales)
summary(model)
```

```
##
## Call:
## lm(formula = price ~ sqft, data = MidWestSales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -239405  -39840   -7641   23515  388362
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -81432.946  11551.846  -7.049 5.74e-12 ***
## sqft           158.950      4.875  32.605  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 79120 on 520 degrees of freedom
## Multiple R-squared:  0.6715, Adjusted R-squared:  0.6709
## F-statistic:  1063 on 1 and 520 DF,  p-value: < 2.2e-16
```

```
n = dim(MidWestSales)[1]
r = cor(MidWestSales$price, MidWestSales$sqft)
Rsquared = summary(model)$r.squared
t = (r*sqrt(n-2))/(sqrt(1-Rsquared))
p = 2*(1-pt(abs(t),n-2))
```

Null hypothesis: B1 = 0, Alternate hypothesis: B1 != 0. Test statistic: t = 32.605, Two sided p-value = 2.2e-16. Since the p-value is essentially 0, which is less that the a = 0.05 level of significance. Therefore, we reject the null hypothesis and can conclude that sqft has a linear relationship with price (alternate hypothesis).

3d) Null hypothesis: B1 = 0, Alternate hypothesis: B1 > 0. For one-sided p-value we divide it by 2, 2.2e-16/2, which is also essentially 0. B1 > 0 and the t* is also positive. Therefore, we reject the null hypothesis and conclude that sqft has a positive linear relationship with price (alternate hypothesis).

3e)

```
predict(model, list(sqft=2000),interval= "c")
```

```
##        fit      lwr      upr
## 1 236467.5 229220.7 243714.4
```

We are 95% confident that the interval (229220.7, 243714.4) contains the true mean of the response (sale price) at X=2000 (2000 sqft).

3f)

```
predict(model, list(sqft=2000),interval= "p")
```

```
##        fit      lwr      upr
## 1 236467.5 80858.85 392076.2
```

We are 95% confident that the interval (80858.85, 392076.2) contains the true response (sale price) at X=2000 (2000 sqft).

3g)

```
predict(model, list(sqft=2000),interval= "p", level = 0.90)
```

```
##        fit    lwr    upr
## 1 236467.5 105948 366987
```

3h) No, it wouldn't make sense as 8500 is greater than the maximum value for sqft. As the model is not fit for that large of a value, predicting the price of a 8500 sqft house may lead to inaccuracy.

3i)

```
predict(model, list(sqft=2000),interval= "p", level = 1)
```

```
##        fit  lwr upr
## 1 236467.5 -Inf Inf
```

The lower and upper limit becomes -infinite and infinite respectively. This is because a 100% confidence means that there is no doubt at all that we will get the same results if the study is repeated.

3j) The estimate of the sigma(e) is 79120, which means that the standard deviation for the approximate distribution of Y is 79120.

3k) Yes it dsigma(e)s make sense for both these houses to have the same estimate of sigma since the standard deviation (or variance) of the Y values is constant for all values of X in the range of the data.

4)

```
skincancer = read.table("/Users/virajvijaywargiya/Downloads/skincancer.txt", fill = TRUE, header = TRUE
NewModel = lm(formula = Mort ~ Lat, data = skincancer)
summary(NewModel)
```

```
##
## Call:
## lm(formula = Mort ~ Lat, data = skincancer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.972 -13.185   0.972  12.006  43.938
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 389.1894    23.8123   16.34  < 2e-16 ***
## Lat          -5.9776     0.5984   -9.99 3.31e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.12 on 47 degrees of freedom
## Multiple R-squared:  0.6798, Adjusted R-squared:  0.673
## F-statistic:  99.8 on 1 and 47 DF,  p-value: 3.309e-13
```

4a)H0 : B1 = 0 vs Ha : B1 6= 0. Test statistic is t* = -9.99. Two sided p-value is 3e-13 which is essentially 0. Therefore, reject the null and conclude evidence for B1 > 0.

4b)

```
predict(NewModel, list(Lat=40),interval= "c", level = 0.99)
```

```
##        fit      lwr     upr
## 1 150.0839 142.7148 157.453
```

The 99% confidence interval for the true mean of the response variable when Lat=40 is (142.7148, 157.453). This means we are 99% confident that the true average mortality at 40 Lat is within that range.

4c)

```
predict(NewModel, list(Lat=40),interval= "p", level = 0.99)
```

```
##        fit      lwr      upr
## 1 150.0839 98.24214 201.9257
```

The 99% confidence interval for the true mean of the response variable when Lat=40 is (98.24214, 201.9257). This means we are 99% confident that the true mortality at 40 lat is within that range.

4d) Both use the same estimate of their respective parameters, uY for mean and Y for true value.

4e) The prediction interval is wider than the confidence interval because it accounts for the error term to the true mean. This will always result in greater variances.

5)5a) Narrower

5b) Narrower

5c) Narrower

5d) Stays the same

6) R-squared ~ 0.73, which means that around 73% of the variance in Y is explained/predicted by X.

7)

```
pulse = read.table("/Users/virajvijaywargiya/Downloads/Pulse.txt", fill = TRUE, header = TRUE)
pulse$Smoker = ifelse(pulse$Smoke==1, "Yes", "No")
view(pulse)
```

7a)

```
pulsemodel = lm(formula = Rest ~ Smoker, data = pulse)
summary(pulsemodel)
```

```
##
## Call:
## lm(formula = Rest ~ Smoker, data = pulse)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.791  -6.041  -0.791   6.209  38.209
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   67.791      0.686  98.826   <2e-16 ***
## SmokerYes      4.978      2.049   2.429   0.0159 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 9.845 on 230 degrees of freedom
## Multiple R-squared:  0.02502,    Adjusted R-squared:  0.02078
## F-statistic: 5.902 on 1 and 230 DF,  p-value: 0.01589
```

Null hypothesis: B1 = 0, Alternate hypothesis: B1 != 0. Test statistic: t-value = 2.429, F-statistic = 5.902. p-value = 0.01589. Therefore, we reject the null hypothesis and conclude that the csigma(e)fficient on X2 (the smoking status) is not 0 (alternate hypothesis). This also means that there is a relationship between resting pulse rate and smoking status.

7b)

```
t.test(pulse$Rest[pulse$Smoker=="Yes"], pulse$Rest[pulse$Smoker=="No"], var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  pulse$Rest[pulse$Smoker == "Yes"] and pulse$Rest[pulse$Smoker == "No"]
## t = 2.4294, df = 230, p-value = 0.01589
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.9405909 9.0153463
## sample estimates:
## mean of x mean of y
##  72.76923  67.79126
```

Having done the two sample t-test, the t-value 2.429 with degrees of freedom 230, and the p-value 0.01589 are the same compared to the test from part a.

7c) Yi = B0 + B1X1 + B2X2 + e

7d) With respect to the effect of weight on resting pulse rate, not having an interaction term means that weight dsigma(e)s not have an effect on smoking status.

7e)

```
summary(lm(formula = Rest ~ Wgt + Smoker, data = pulse))
```

```
##
## Call:
## lm(formula = Rest ~ Wgt + Smoker, data = pulse)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.872  -6.207  -0.719   5.794  37.128
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 78.24692    3.22061  24.296  < 2e-16 ***
## Wgt         -0.06697    0.02017  -3.319  0.00105 **
## SmokerYes    6.04288    2.03136   2.975  0.00325 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.638 on 229 degrees of freedom
## Multiple R-squared:  0.06978,    Adjusted R-squared:  0.06165
## F-statistic: 8.589 on 2 and 229 DF,  p-value: 0.0002531
```

Equation: Y = 78.25 - 0.067 X1 + 6.04X2 Csigma(e)fficient of determination (multiple R-squared value) = 0.06978, estimate of sigma(e) (RSE) = 9.638.

7f) 232

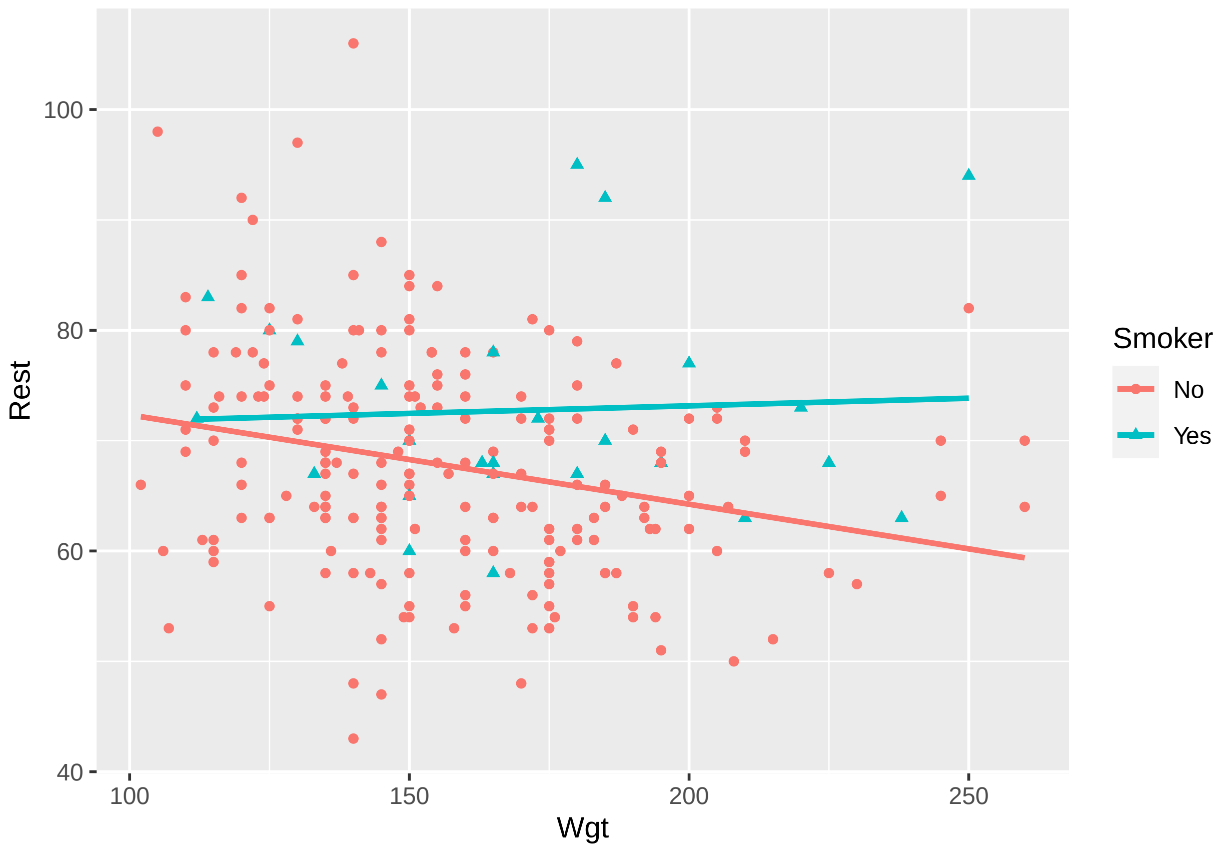
7g) Null hypothesis: B2 = 0, Alternate hypothesis: B2 != 0. Test statistic: t-value = 2.975. p-value = 0.00325. Therefore, we reject the null hypothesis and conclude that the csigma(e)fficient on smoking yes (X2) in not 0, which means that with respect of the effect of weight on resting pulse rate, there is a relation between resting pulse and smoking status.

7h) No, we cannot conclude that smoking causes lower resting pulse rates because it is an observational study and the cofounding variables are not in control. Also, smoking status and resting pulse rates have a positive relation.

7i)

```
library(ggplot2)
pulse$Smoker = ifelse(pulse$Smoke==1, "Yes", "No")
ggplot(pulse, aes(Wgt, Rest, color = Smoker, shape=Smoker)) + geom_point()+geom_smooth(method="lm", se=
```

```
## `geom_smooth()` using formula 'y ~ x'
```



7j) $Yi = B0 + B1X1 + B2X2 + (BX1X2)+ e$

7k)

```
summary(lm(formula = Rest ~ Wgt + Smoker + Wgt*Smoker, data = pulse))
```

```
##
## Call:
## lm(formula = Rest ~ Wgt + Smoker + Wgt * Smoker, data = pulse)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.097  -6.182  -0.752   5.832  36.903
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     80.43031    3.46360  23.222  < 2e-16 ***
## Wgt             -0.08095    0.02177  -3.719 0.000252 ***
## SmokerYes      -10.03147    9.82355  -1.021 0.308258
## Wgt:SmokerYes    0.09473    0.05665   1.672 0.095863 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.6 on 228 degrees of freedom
## Multiple R-squared:  0.08105,    Adjusted R-squared:  0.06896
## F-statistic: 6.703 on 3 and 228 DF,  p-value: 0.0002356
```

Regression equation: Y = 80.43 - 0.081 X1 - 10.031 X2 + (0.095 X1X2)

7l) R-squared value = 0.08105, sigma(e) = 9.6. Compared to part e, there is a slight increase in the R-squared value, but the sigma(e) remains about the same.

7m) Null hypothesis: B3 = 0, Alternate hypothesis: B3 != 0. Test statistic: t-value = 1.672. p-value = 0.095863. Therefore, we fail to reject the null hypothesis, and can conclude that the effect of weight on resting pulse rate can differ for smokers and non-smokers.

7n) No, it is not a good fitting model based on the R-squared value because a value of 0.07 means that only 7% of the variance in Y is explained by the eplanatory variables Xi.

7o) Exercising increases blood flow to the hear and also increases the pulse rates. This could be a cofounder in this study as the the response variable pulse rate dsigma(e)s get affected by Exercise amount.

7p) This can be done by checking if the r-squared value increases when changing X1. Null hypothesis: X1 > 0, Alternate hypothesis: X1 <= 0.