HW2_STAT210P_MinhCao

2025-01-31

1. Say a regression equation if fit to data, and the correlation coefficient estimate ,R ,between X and Y is 0.5.

State if true or false.

a.The slope of the regression line is 0.5: False

b.The regression model with X explains 50% of the variation in Y: False

c.25% of the variation in Y is explained by X: True

d.50% of the variation in Y is explained by X: False (R^2 = 0.25: coefficient of determination) (R = 0.5: Correlation Coefficient)

e. X is positively associated with Y: True

f.Even if X and Y have a non linear relationship, the value of 0.5 can be used to measure the association between X and Y.: False

2 Suppose a simple linear model is fit to predict Y = weight in kilograms using X= height in centimeters of an adult. But say a new simple linear model is fit using Y = height in centimeters and X= weight in kilograms (that is to say Y and X have reversed).State whether each of the following would be the same for this new model as it was for the original model, or it would be different and explain in a sentence or two.

a. The value of R.: Stay the same as the formula stay the same

b. The value of R2: Stay the same as value of R stay the same

c. The estimate of B1: Different, In the original model, B1 represents the change in Weight per unit change in Height. In the new model, it represents the change in Height per unit change in Weight.

d. The estimate of B0: Different. The intercept B0 represents the expected value of the dependent variable when the independent variable is zero

e. The test statistic to test if the correlation between explanatory and response is equal to 0.: Same because correlation stays the same

3

```
mid_path = "C:/Users/caoqu/OneDrive/Desktop/UCI/STUDY/WINTER_25/DATA_210P_Statistical_Methods_I/data/Mi
Midwest = read.table(mid_path, sep= "", header = FALSE)
head(Midwest)
```

```
##   V1     V2   V3 V4 V5 V6 V7 V8   V9 V10 V11   V12 V13
## 1  1 360000 3032  4  4  1  2  0 1972   2   1 22221   0
## 2  2 340000 2058  4  2  1  2  0 1976   2   1 22912   0
## 3  3 250000 1780  4  3  1  2  0 1980   2   1 21345   0
## 4  4 205500 1638  4  2  1  2  0 1963   2   1 17342   0
## 5  5 275500 2196  4  3  1  2  0 1968   2   7 21786   0
## 6  6 248000 1966  4  3  1  5  1 1972   2   1 18902   0
```

```
names(Midwest)=c("id","price","sqft","bed","bath","ac","garage","pool","year","quality",
"style","lot","hwy")
```

a. Fit a linear model where the square footage of the house is used to predict the sale price. (X=sqft and Y =price). Write the estimated regression equation

```
model1 = lm(price ~ sqft, data = Midwest)
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ sqft, data = Midwest)
##
## 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
```

- Estimated regression equation: Price = -81432.946 + 158.950*sqft

b. Interpret the estimate of the slope.

When average sqft (X) increases by 1 unit, we expect the price of the house to increase by $158.950

c. Test whether square footage has a significant linear relationship with price. Use alpha = 0.05 level of significance. Write the null and alternative hypothesis, state the test statistic and p-value, and make a conclusion

R^2 = 0.6715 and B1 = 158.950

```
# Therefore R =

R = 0.6715**0.5
R
```

```
## [1] 0.819451
```

- H0: p = 0
- Ha: p != 0
- Test statistics t* =

```
# Test statistics t*
df = 520
t_star = (R * sqrt(df - 2)) / (sqrt(1 - R**2))
t_star
```

```
## [1] 32.5402
```

```
# Finding p-value

p_value = 2*pt(q = t_star, df = 518, lower.tail = FALSE)
p_value
```

```
## [1] 2.589982e-127
```

Using alpha = 0.05 and p_value = 2.589982e-127, we reject null hypothesis and accept alternative hypothesis. We conclude that square footage has a significant linear relationship with price

d. Now test whether square footage has a significant positive linear relationship with price. Use alpha = 0.05 level of significance. Write the null and alternative hypothesis, state the test statistic and p-value, and make a conclusion.

Same as c, we reject null hypothesis and accept alternate hypothesis, and since B1 > 0, we conclude that square footage has a significant positive linear relationship with price

e.Find and interpret a 95% confidence interval for the mean price when sqft=2000

```
# 95% confidence interval

predict(model1, list(sqft = 2000), interval = "c" )
```

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

f.Find and interpret a 95% prediction interval for the price when sqft=2000

```r
# 95% prediction interval
predict(model1, list(sqft = 2000), interval = "p" )
```

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

g. What would happen to the interval from part f. if the confidence level is decreased to 90%? Explain in a sentence or two

The interval will be smaller as the multiplier become smaller

h.Would it make sense to predict the sale price of a house that is 8500 square feet?

Yes

4.This question will use the skin cancer data set that is on the class website. Say it is of interest if latitude is predictive of mortality rate due to skin cancer. Fit a simple linear model where X=Lat and Y =Mort.

```r
skin_path = "C:/Users/caoqu/OneDrive/Desktop/UCI/STUDY/WINTER_25/DATA_210P_Statistical_Methods_I/data/sk
skin = read.table(skin_path, sep= "", header = TRUE)
head(skin)
```

```
##         State  Lat Mort Ocean  Long
## 1     Alabama 33.0  219     1  87.0
## 2     Arizona 34.5  160     0 112.0
## 3    Arkansas 35.0  170     0  92.5
## 4  California 37.5  182     1 119.5
## 5    Colorado 39.0  149     0 105.5
## 6 Connecticut 41.8  159     1  72.8
```

```r
model2 = lm(Mort ~ Lat, data = skin)
summary(model2)
```

```
##
## Call:
## lm(formula = Mort ~ Lat, data = skin)
##
## 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
```

a. Using a alpha = 0.05 significance level, conduct a formal statistical test of whether latitude has a linear association with the mortality.

R^2 = 0.6715 and B1 = 158.950

```
# Therefore R =

R = 0.6798**0.5
R
```

```
## [1] 0.8244998
```

- H0: p = 0
- Ha: p != 0
- Test statistics t* =

```
# Test statistics t*
df = 47
t_star = (R * sqrt(df - 2)) / (sqrt(1 - R**2))
t_star
```

```
## [1] 9.774311
```

```
# Finding p-value

p_value = 2*pt(q = t_star, df = 45, lower.tail = FALSE)
p_value
```

```
## [1] 1.057931e-12
```

Using alpha = 0.05 and p_value = 1.057931e-12, we reject null hypothesis and accept alternative hypothesis. We conclude that latitude has a linear association with the mortality.

b. Find and interpret a 99% confidence interval for the mean mortality rate when Lat=40.

```
# 99% confidence interval
predict(model2, list(Lat = 40), interval = "c" )
```

```
##        fit      lwr      upr
## 1 150.0839 144.5617 155.6061
```

c. Find and interpret a 99% prediction interval for the mortality rate when Lat=40.

```
# 99% prediction interval
predict(model2, list(Lat = 40), interval = "p" )
```

```
##        fit     lwr      upr
## 1 150.0839 111.235 188.9329
```

d. What can you say about the center of the confidence interval and prediction interval. Is it the same?

Yes they are the same. Both intervals are centered at the same point—the predicted value from your model

e. How does the width of the confidence interval compare to the prediction interval. Explain in a sentence or two.

The prediction interval is wilder than the confidence interval because it includes the extra uncertainty from the variability of individual observations, not just the uncertainty in estimating the mean

5 Remember the form of the prediction interval. The standard error of the prediction at a value of X = Xp was SE(Y^p)

For each part, state what will happen to the prediction interval (stay the same, be wider, or be narrower).

a. The sample size is increased.: This make the prediction interval narrower
b. If Xp gets closer to X bar, the average of the X covariate: This make the prediction interval narrower
c. Wider
d. Stay the same