Module 3 Assesment

BTRY 6020

Question 1 (2 pts)

Suppose I believe my data is generated by the following model:

$$Y_i = b_0 + b_1 X_{i,1} + b_2 X_{i,2} + b_3 X_{i,3} + b_4 X_{i,4} + \varepsilon_i.$$

I want to test the null hypothesis that $X_{i,2}$ is not associated with Y_i after adjusting for $X_{i,1}$, $X_{i,3}$, and $X_{i,4}$. The alternative hypothesis is that there is some association between $X_{i,2}$ and Y_i, even after adjusting for the other covariates. What is the null hypothesis and the alternative hypothesis:

Answer

$H_0: \ b_2 = 0$
$H_A: \ b_2 \neq 0$

Question 2 (2 pts)

Suppose I gather 35 observations and fit the model specified above. Given the output below, calculate the t-statistic for testing the hypothesis. Round this answer to two digits after the decimal.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.971	0.171	11.533	0
X1	0.861	0.212	4.063	0
X2	0.373	0.194	???	???
X3	1.078	0.188	5.738	0
X4	-0.057	0.221	-0.259	0.798

Answer

$t = \frac{\hat{b}_2}{\hat{SE}(\hat{b}_2)} = \frac{0.373}{0.194} = 1.923$

Question 3 (2 pts)

Suppose you are interested in testing the null hypothesis $H_0 : b_2 = .5$. Given the table above, calculate the t-statistic for testing this null hypothesis. Round this answer to two digits after the decimal.

Answer

$t = \frac{\hat{b}_2 - b_2^{(0)}}{\hat{SE}(\hat{b}_2)} = \frac{0.373 - 0.5}{0.194} = -0.655$

Question 4 (2 pts)

You can use the `qt` function to get the cut-off from a T distribution. Specifically, the code below gets the cutoff so that the area to right of that cut-off is alpha / 2 for a T distribution with z degrees of freedom. Calculate the p-value for the table in Question 2, and also for the null hypothesis in Question 3.

```
qt(alpha / 2, df = z, lower = F)
```

Problem 2: $p = P(|t_{n-p-1}| \geq 1.923) = 2 \cdot P(t_{n-p-1} < -1.923) = 2 \times$ `pt(-1.923,35-4-1)` $= 0.064$
Problem 3: $p = P(|t_{n-p-1}| \geq 0.655) = 2 \cdot P(t_{n-p-1} < -0.655) = 2 \times$ `pt(-0.655,35-4-1)` $= 0.517$

Question 5 (1 pt)

Calculate a 90% confidence interval for the coefficient of X_1.

Answer

$\hat{b}_1 \pm \hat{SE}\left(\hat{b}_1\right) \cdot t_{n-p-1}^{(\alpha/2)}$
$0.861 \pm (0.212) t_{35-4-1}^{(0.05)}$
$0.861 \pm (0.212)(1.697)$
$[0.501, 1.221]$

Question 6 (2 pts)

Consider two worlds. In both, you are interested in testing the null hypothesis that $H_0 : b_1 = 0$ vs $H_A : b_1 \neq 0$. In the first setting $b_1 = 1$ and in the second setting $b_1 = 2$. If all other things are equal, in which setting do you have more power to reject the null hypothesis. Give a brief explanation of why?

Answer

In both worlds, presuming all model assumptions are met, the estimates $\hat{b}_1$ are distributed normally around their true value. Since 2 is farther away from 0 than 1 is, the sampling distribution of $\frac{\hat{b}_1}{\hat{\text{var}}(\hat{b}_1)}$ will have more probability mass in the rejection region when $b_1 = 2$ when compared to the sampling distribution when $b_1 = 1$, assuming the variance of the residuals is the same in both settings.

Question 7 (2 pts)

Suppose you are interested in testing the null hypothesis that $H_0 : b_1 = 0$ vs $H_A : b_1 \neq 0$. However, the true $b_1 = 1$. Suppose you are deciding to test the null hypothesis with either a $\alpha = .05$ or $\alpha = .1$ level test. All other things are equal, in which test would have more power to reject the null hypothesis. Give a brief explanation of why?

Answer

With a larger α, we are tolerating a higher false-positive (type 1) error rate. Thus, our rejection region is larger and we would reject the null hypothesis for smaller (in absolute value) t-statistics. This means we will have higher power to reject the null hypothesis when it is false.

Housing Data

Recall the housing data that we've been considering in lecture. We can load the data using the following code:

```
fileName <- url("https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lectureData/estate.csv")
housing_data <- read.csv(fileName)

head(housing_data)
```

```
##   id  price area bed bath  ac garage pool year quality style   lot highway
## 1  1 360000 3032   4    4 yes      2   no 1972  medium     1 22221      no
## 2  2 340000 2058   4    2 yes      2   no 1976  medium     1 22912      no
## 3  3 250000 1780   4    3 yes      2   no 1980  medium     1 21345      no
## 4  4 205500 1638   4    2 yes      2   no 1963  medium     1 17342      no
## 5  5 275500 2196   4    3 yes      2   no 1968  medium     7 21786      no
## 6  6 248000 1966   4    3 yes      5  yes 1972  medium     1 18902      no
```

There are 522 observations with the following variables:

- price: in 2002 dollars
- area: Square footage
- bed: number of bedrooms
- bath: number of bathrooms
- ac: central AC (yes/no)
- garage: number of garage spaces
- pool: yes/no
- year: year of construction
- quality: high/medium/low
- home style: coded 1 through 7
- lot size: sq ft
- highway: near a highway (yes/no)

There is no age data in the table, but we can compute it on our own from the year variable

```
housing_data$age <- 2002 - housing_data$year
```

Question 8 (3 pts)

Let log(*price*) be the dependent variable. Suppose we are interested in the association of log(*price*) with the lot size, after conditioning for the area, age, and number of bedrooms. Estimate the linear coefficient of interest and give an interpretation of the point estimate. Form a 95% confidence interval for the coefficient of interest.

Answer

```
lmod8 <- lm(log(price)~lot+area+age+bed,data=housing_data)
coef(lmod8)
```

```
##   (Intercept)           lot          area           age           bed
##  1.160456e+01  6.086119e-06  4.123965e-04 -7.318639e-03  1.823439e-03
```

```
confint(lmod8)
```

```
##                     2.5 %        97.5 %
## (Intercept)  1.150955e+01  1.169956e+01
## lot          4.563026e-06  7.609211e-06
## area         3.805950e-04  4.441980e-04
## age         -8.428175e-03 -6.209102e-03
## bed         -1.867326e-02  2.232014e-02
```

The point estimate for the coefficient of interest is $6.086 \cdot 10^{-6}$. This means that when comparing two houses which differ in lot size by 1 square foot, the price will go up by approximately 0.0006086 percent, holding area, age, and number of bedrooms constant. (Note that $e^{\hat{b}_1} - 1$ is very close to $\hat{b}_1$ because $e^x - 1 \approx x$ near $x = 0$) We are 95% confident that b_1 lies between $2.84 \cdot 10^{-6}$ and $9.33 \cdot 10^{-6}$.

Question 9 (3 pts)

Let $\log(price)$ be the dependent variable. Suppose we are interested in the association of $\log(price)$ with the number of bedrooms, after conditioning for the $\log(area)$, $\log(lot)$, and age. Conduct a hypothesis test with level $\alpha = .05$ for the null hypothesis that bedrooms is not associated with $\log(price)$ after conditioning for $\log(area)$, $\log(lot)$, and age. What is the resulting t statistic? What is the result of the hypothesis test?

Answer

```
lmod9 <- lm(log(price)~bed+log(area)+log(lot)+age,data=housing_data)
summary(lmod9)
```

```
##
## Call:
## lm(formula = log(price) ~ bed + log(area) + log(lot) + age, data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67070 -0.11466 -0.00898  0.10479  0.86152
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.203514   0.321595   9.961  < 2e-16 ***
## bed         -0.005632   0.010296  -0.547    0.585
## log(area)    1.031987   0.039419  26.180  < 2e-16 ***
## log(lot)     0.156110   0.021647   7.212 1.98e-12 ***
## age         -0.006745   0.000554 -12.176  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1957 on 517 degrees of freedom
## Multiple R-squared:  0.796,  Adjusted R-squared:  0.7944
## F-statistic: 504.2 on 4 and 517 DF,  p-value: < 2.2e-16
```

The t-statistic for this hypothesis test is $t = -0.547$. The resulting p-value is 0.585 which indicates a lack of evidence that the number of beds is associated with the log price. We thus fail to reject the null hypothesis that $b_{bed} = 0$.

Question 10 (3 pts)

Let $\log(price)$ be the dependent variable. Suppose we are interested in the association of $\log(price)$ with quality of the house, after conditioning for the $\log(area)$, age, and number of bedrooms. Conduct a hypothesis test with level $\alpha = .05$ for the null hypothesis that quality is not associated with $\log(price)$ after conditioning for the area, age, and number of bedrooms. What is the resulting statistic? What is the result of the hypothesis test?

Answer

Since ‘quality‘ is a categorical variable, we must test the coefficients of all the dummy variables at the same time. For this, we conduct a regression F-test:

```
lmod10a <- lm(log(price)~quality + log(area) + age + bed, data=housing_data)
lmod10b <- lm(log(price)~          log(area) + age + bed, data=housing_data)
anova(lmod10a,lmod10b)
```

```
## Analysis of Variance Table
##
## Model 1: log(price) ~ quality + log(area) + age + bed
## Model 2: log(price) ~ log(area) + age + bed
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    516 17.393
## 2    518 21.800 -2   -4.4073 65.377 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With an F statistic of 65.377 on 516 and 518 degress of freedom, we find a p-value smaller than $2 \cdot 10^{-16}$. Thus we reject the null hypothesis and conclude that the quality of the house is associated with the price of the house.