hw2_yw3204

NAME: Yuhao Wang, UNI: yw3204

9/28/2018

2.1

a.

Since 0 is not in the confidence interval, the conclusion is warranted by the duality of confidence interval and hypothesis testing.

We see the student is estimating the 95 percent confidence interval, and thus the level of significance is 5%.

b.

It is right that dollar sales can't be negative. And we can improve our lower limit to 0.

2.4

a.

```r
gpa <- read.table("CH01PR19.txt", header = FALSE)
names(gpa) <- c("Y", "X")
n <- nrow(gpa)
reg <- lm(Y ~ X, data = gpa)
summ <- summary(reg)
sd_1 <- summ$coefficients[2, 2]
mu_1 <- summ$coefficients[2, 1]
c1 <- mu_1 + qt(0.005, n-2) * sd_1
c2 <- mu_1 + qt(0.995, n-2) * sd_1
CI <- c(c1, c2)
CI
```

```
## [1] 0.005385614 0.072268640
```

```r
# the following function can be called to calculate the CI directly
# confint(reg, "X", 0.99)
```

The 99 percent confidence interval of β_1 is listed above. It doesn't include 0. Since confidence interval and hypothesis testing are two sides of the same coin, the director is actually concerned about whether the ACT score has an impact on GPA.

b.

We construct the t^* like below: $t^* = \frac{\hat{\beta_1}}{se(\hat{\beta_1})}$ in which, $se(\hat{\beta_1}) = \frac{\hat{\sigma}}{\Sigma(x_i - \bar{x})^2}$.

Thus, the null hypothesis is: $H_0 : t^* = 0$, and the alternative is: $H_1 \neq 0$.

Since the significance level is given 0.01, the decision rule should be: if $t^* > t(0.005, 118)$, in which $t(0.005, 118)$ is the 0.005 quantile of t distribution of df 118, we approve the alternative.

```
t_star <- mu_1 / sd_1
t_star > qt(0.995, n-2)
```

```
## [1] TRUE
```

Based on the result above, we take the alternative that states there is a linear relationship between ACT score and gpa.

c.

```
2*(1-pt(t_star, n-2))
```

```
## [1] 0.002916604
```

The P-value is given above. Because it is samller than 0.01, it is consistent with the conclusion in (b).

2.7

a.

```
plastic <- read.table("CH01PR22.txt")
names(plastic) <- c("Y", "X")
n_1 <- nrow(plastic)
reg_1 <- lm(Y ~ X, plastic)
confint(reg_1, "X", 0.99)
```

```
##      0.5 %   99.5 %
## X 1.765287 2.303463
```

The 99 percent confidence interval is calculated above. It means there is a 99 percent of probability that the change in the mean hardness will fall into the interval.

b.

According to the question, the null hepothesis is: $H_0 : \beta_1 = 2$, and the alternative is: $H_1 : \beta_1 \neq 2$.

Comparing the result from (a), since 2 is in the confidence interval, we approve the null.

To calculate the P-value, we construct the statistic below: $t = \frac{\hat{\beta_1}-2}{se(\hat{\beta_1})}$

```
summ_1 <- summary(reg_1)
b1 <- summ_1$coefficients[2, 1]
sd_b1 <- summ_1$coefficients[2, 2]
t_1 <- (b1 - 2) / sd_b1
(1-pt(t_1, n_1-2))*2
```

```
## [1] 0.7094445
```

The P-value is provided above which further consolidates our conclusion.

c.

We accept the null hypotehsis when P-value is smaller than 0.01. It is equivelent to

$|\frac{\hat{\beta}_1-2}{se(\hat{\beta}_1)}| > |t(0.005, 14)|$.

Thus, the range of $\hat{\beta}_1$ is:

$\hat{\beta}_1 < 2 - 0.1 * |t(0.005, 14)|$ or $\hat{\beta}_1 > 2 + 0.1 * |t(0.005, 14)|$.

Given that the true $\beta_1 = 2.3$, we calcule the conditional probability:

$\mathbb{P}(\{\hat{\beta}_1 < 2 - 0.1 * |t(0.005, 14)|\} \cup \{\hat{\beta}_1 > 2 + 0.1 * |t(0.005, 14)|\} | \beta_1 = 2.3)$

```
p <- pt(-0.3+qt(0.005, 14)*0.1, 14) + 1 - pt(-0.3-qt(0.005, 14)*0.1, 14)
1-p
```

```
## [1] 0.219295
```

Therefore, the power of test is 0.219295.

2.12

Notice that $\sigma^2\{pred\} = \sigma^2 + \sigma^2\{\hat{Y_h}\}$.

Because the first term has nothing to do with n and $\sigma^2\{\hat{Y_h}\} \geq 0$, no matter how large n is, $\sigma^2\{pred\}$ is always greter than or equal to σ^2 and won't converge to 0.

But for $\sigma^2\{\hat{Y_h}\}$, it will converge to 0.

Accoridng to the formula: $\sigma^2\{\hat{Y_h}\} = \sigma^2[\frac{1}{n} + \frac{X_h-\bar{X})^2}{\Sigma(X_i-\bar{X})^2}]$, the first term in the bracket $\frac{1}{n}$ is undoubtedly converge to 0.

For the latter part, because:

$\bar{X} \xrightarrow{P} \mathbb{E}(X)$ and $\frac{1}{n-1}\Sigma(X_i - \bar{X})^2 \xrightarrow{P} Var(X)$,

we have:

$\bar{X} = O_p(1)$ and $\frac{1}{n-1}\Sigma(X_i - \bar{X})^2 = O_p(1)$.

Thus, given a fixed value X_h and by the Continuous Mapping Theorem, we have: $\frac{X_h-\bar{X})^2}{\frac{1}{n-1}\Sigma(X_i-\bar{X})^2} = O_p(1)$, which means it is bounded in probability.

Therefore, $\frac{(X_h-\bar{X})^2}{\Sigma(X_i-\bar{X})^2} = \frac{1}{n-1} * \frac{X_h-\bar{X}}{\frac{1}{n-1}\Sigma(X_i-\bar{X})^2} \to 0$

The diffenrnce implied that we can not erase the uncertainty raised by the error term in the model when predicting a new observation. However, we can always decrease the uncertainty when predicing the true mean by increasing the sample size.

2.13

a.

```
CI1 <- c()
mu_x <- summ$coefficients[1, 1] + summ$coefficients[2, 1] * 28
CI1[1] <- mu_x+qnorm(0.025)*summ$sigma*sqrt(1/n + (28-mean(gpa$X))^2 / ((n-1) * var(gpa$X)))
CI1[2] <- mu_x-qnorm(0.025)*summ$sigma*sqrt(1/n + (28-mean(gpa$X))^2 / ((n-1) * var(gpa$X)))
CI1
```

```
## [1] 3.062818 3.339599
```

```r
# we can also call the function below:
# predict(reg, X6, se.fit = TRUE, interval = "confidence", level = 0.95)
```

The interval is calculated above and it means that the mean GPA at point ACT score equals 28 will fall into the interval with probability 95%.

b.

```r
CI2 <- c()
CI2[1] <- mu_x+qnorm(0.025)*summ$sigma*sqrt(1 + 1/n + (28-mean(gpa$X))^2 / ((n-1) * var(gpa$X)))
CI2[2] <- mu_x-qnorm(0.025)*summ$sigma*sqrt(1 + 1/n + (28-mean(gpa$X))^2 / ((n-1) * var(gpa$X)))
CI2
```

```
## [1] 1.972090 4.430327
```

```r
# predict(reg, X6, se.fit = TRUE, interval = "prediction", level = 0.95)
```

The interval is provided above and it stands for that Mary's GPA will fall into the interval with probability 95%.

c.

Yes, it is wider. Unlike predicting the expected value, the prediction for new observation is influenced by one more error term. Thus, it is more uncertain and should require a wider confidence interval.

d.

```r
CI3 <- c()
W <- sqrt(2 * qf(0.95, 2, n-2))
CI3[1] <- mu_x-W*summ$sigma*sqrt(1/n + (28-mean(gpa$X))^2 / ((n-1) * var(gpa$X)))
CI3[2] <- mu_x+W*summ$sigma*sqrt(1/n + (28-mean(gpa$X))^2 / ((n-1) * var(gpa$X)))
CI3
```

```
## [1] 3.026159 3.376258
```

The interval is calculated above and it is wider than that of (a), because the confidence band must encompass the entire regression line.

2.51

First, we prove: $\mathbb{E}(b_1) = \beta_1$

$\mathbb{E}(b_1) =$

$= \mathbb{E}(\frac{\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{\Sigma(X_I-\bar{X})^2})$

$= \mathbb{E}(\frac{\Sigma(X_i-\bar{X})Y_i}{\Sigma(X_I-\bar{X})^2}) - \mathbb{E}(\frac{\Sigma(X_i-\bar{X})\bar{Y}}{\Sigma(X_I-\bar{X})^2})$

$= \frac{\Sigma(X_i-\bar{X})\mathbb{E}(Y_i)}{\Sigma(X_I-\bar{X})^2}$

$= \frac{\Sigma(X_i-\bar{X})(\beta_0+\beta_1 * X_i)}{\Sigma(X_I-\bar{X})^2}$

$= \beta_1 * \frac{\Sigma(X_i-\bar{X})X_i}{\Sigma(X_I-\bar{X})^2}$

$= \beta_1 * \frac{\Sigma(X_i - \bar{X})(X_i - \bar{X})}{\Sigma(X_I - \bar{X})^2}$

$= \beta_1$

Then, we have:

$\mathbb{E}(b_0)$

$= \mathbb{E}(\bar{Y}) - \mathbb{E}(b_1) * \bar{X}$

$= \beta_0 + \beta_1 * \bar{X} - \beta_1 * \bar{X}$

$= \beta_0$

Therefore, we proved b_0 is an unbiased estimator of β_0.

2.52

By definition, we have: $\sigma^2\{b_0\} = Var(\bar{Y} - b_1 * \bar{X})$

$= Var(\bar{Y}) + \bar{X}^2 * Var(b_1) - 2 * \bar{X} * Cov(\bar{Y}, b_1))$

$= Var(\bar{Y}) + \bar{X}^2 * Var(b_1)$

$= \frac{\sigma^2}{n} + \bar{X}^2 * \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2}$

$= \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2})$

Observing the formula of $\sigma^2\{\hat{Y}_h\} = \sigma^2(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2})$, we find $\sigma^2\{b_0\}$ is a special case of $\sigma^2\{\hat{Y}_h\}$ in which $X_h = 0$.

2.63

```
CDI <- read.table("APPENC02.txt")

CIs <- list()

for(i in 1:4) {
  ind = CDI$V17 == i;
  lm <- lm(V15 ~ V12, data = CDI[ind, ]);
  CIs[[i]] <- confint(lm, "V12", 0.9)
}

CIs[[1]]
```

```
##           5 %  95 %
## V12 460.5177 583.8
```

```
CIs[[2]]
```

```
##           5 %    95 %
## V12 193.4858 283.853
```

```
CIs[[3]]
```

```
##           5 %     95 %
## V12 285.7076 375.5158
```

```
CIs[[4]]
```

```
##          5 %     95 %
## V12 364.7585 515.8729
```

The confidence interval for each region is provided above and the regression lines appear to have different slopes.