Homework 2

Name: Quan Yuan UNI: qy2205

2.1

(a) Yes, the conclusion is warranted. The confidence interval for the slope is [0.452886, 1.05721], which not includes 0. So at 5% significance level, the slope is significantly different than 0. (if H0: slope =0, we will reject the null hypothesis)

(b) The result of the regression model highly depends on the data we use, the real relationship may like $y = x^3$ $(x \geq 0)$. But when we only use data with x much bigger than 0, the regression line may have negative intercept. Besides, the value of $x = 0$ is not important since we don't need to consider the sales volume in an area with population equals to 0.

2.4

(a) Since σ^2 is unknown, we could get,

$$\frac{\widehat{\beta_1} - \beta_1}{\frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}} = \frac{\widehat{\beta_1} - \beta_1}{\frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}\frac{\hat{\sigma}}{\sigma}} \sim t(n-2)$$

$$\mathrm{P}\left(\mathrm{t}_{0.005}(\mathrm{n}-2) \leq \frac{\widehat{\beta_1} - \beta_1}{\frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}}) \leq \mathrm{t}_{0.995}(\mathrm{n}-2) \right) = 0.99$$

So, based on the above formula, we could get

$$\beta_1 \in [\widehat{\beta_1} - \mathrm{t}_{0.995}(\mathrm{n}-2)\frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}, \widehat{\beta_1} - \mathrm{t}_{0.005}(\mathrm{n}-2)\frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}]$$

According to the result in 1.19

$$\widehat{\beta_1} = 0.039, \mathrm{t}_{0.995}(118) = 2.618, \mathrm{t}_{0.005}(118) = -2.618$$

$$\hat{\sigma} = 0.623, \sqrt{\sum(x_i - \bar{x})^2} = 48.784$$

Therefore, the confidence interval for β_1 is

$$[0.0054, 0.0723]$$

Does it include zero? No

Why might the director of admissions be interested in whether the confidence interval includes zero? Because if the confidence interval include 0, we cannot reject the null hypothesis test that $\beta_1 = 0$, which means there is no relationship between ACT and GPA.

(b) a = 0.01

First set hypothesis:

$$\mathrm{H0}: \beta_1 = 0$$

$$H1: \beta_1 \neq 0$$

Then we calculate t-statistic

$$t = \frac{\widehat{\beta_1}}{\frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}} \sim t_{n-2}$$

$$t^* = \frac{0.039}{0.623/48.784} = 3.054$$

Since

$$t^* > t(1-a, n-2) = t(0.99, 118) = 2.3584$$

So, we reject H0, there is a linear relationship between ACT and GPA.

(c)

$$p - value = \Pr(|t_{118}| \geq 3.039) = 0.0029 \leq 0.01$$

$p - value$ equals to 0.0029 means the probability of $\beta_1 = 0$ is 0.29% which is very small. So we can reject the null hypothesis.

2.7

(a) For calculating the change, we need to estimate β_1 and also get 99% confidence interval. Since we still don't know the standard deviation, so the confidence interval is

$$\beta_1 \in [\widehat{\beta_1} - \mathrm{t}_{0.995}(\mathrm{n}-2)\frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}, \widehat{\beta_1} - \mathrm{t}_{0.005}(\mathrm{n}-2)\frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}]$$

We use Python to calculate the required value.

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from scipy.stats import t
```

```
# import data
data = pd.read_table('CH01PR22.txt', header = None, sep = ' ')[[2, 5]]
data.columns = ['hardness', 'elapsed']
data.head()
```

	hardness	elapsed
0	199.0	16.0
1	205.0	16.0
2	196.0	16.0
3	200.0	16.0
4	218.0	24.0

```
# linear regression model
lm0 = LinearRegression()
lm0.fit(X = data['elapsed'].values.reshape(-1,1), y = data['hardness'])
# slope
print('slope is', lm0.coef_[0])
# intercept
print('slope is', lm0.intercept_)
# residual
residual = lm0.predict(data['elapsed'].values.reshape(-1,1)) - data['hardness'].values
# sigma
sigma = np.sqrt(sum(residual*residual)/(len(data)-2))
print('sigma is',sigma)
# sqrt(sum(x_i - x_bar)^2)
print('sqrt(sum(x_i - x_bar)^2) is', np.sqrt(sum((data['elapsed'] - np.mean(data['elapsed']))**2)))
# calculate t-statistic
print('t-statistic is', t.ppf(0.995, df = 14))
```

```
slope is 2.034375
slope is 168.60000000000002
sigma is 3.234026680692135
sqrt(sum(x_i - x_bar)^2) is 35.77708763999664
t-statistic is 2.97684273411266
```

After that, we can calculate the confidence interval is [1.77, 2.30], which means when the elapsed time increases by one hour, the probability of the change in the mean hardness within 1.77 to 2.30 is 99%.

(b) Set hypothesis:

$$H_0: \beta_1 = 2, H_1: \beta_1 \neq 2$$

Test Statistics:

$$t = \frac{\widehat{\beta_1} - 2}{\hat{\sigma}/\sqrt{\sum(x_i - \bar{x})^2}} \sim t_{n-2} = t_{14}$$

Observed t:

$$t = \frac{2.0344 - 2}{\frac{3.2340}{35.7771}} = 0.3803$$

$$p - value = Pr(|t_{14}| \geq 0.3803) = 0.7094 \geq 0.01$$

So we accept the null hypothesis.

(c)

$$a = 0.01, \delta = \frac{\Delta mean}{\sigma} = \frac{0.3}{0.1} = 3, df = 118$$

According to the table B.5 in Appendix, $Power = 0.65$

2.12

$$\sigma^2(Pred) = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$$\lim_{n\to\infty} \frac{1}{n} = 0$$

$$\lim_{n\to\infty}\frac{(x_h-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=(x_h-\bar{x})^2$$

$$\lim_{n\to\infty}\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=0, x_i\neq\bar{x}$$

$$\lim_{n\to\infty}\sigma^2(Pred)=\sigma^2$$

$$\lim_{n\to\infty}\sigma^2(Y_h)=0$$

The difference is because the observed values will always have random errors that follow normal distribution.

2.13

(a) The 95% confidence interval of μ_x is

$$\mu_x\in\left[\widehat{\mu_x}-1.96\cdot\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}},\widehat{\mu_x}+1.96\cdot\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}}\right]$$

$$\widehat{\mu_x}=\widehat{\beta_0}+\widehat{\beta_1}x=2.1147+0.0388\times28=3.2011,\mathrm{n}=120$$

$$\hat{\sigma}=0.623,\sqrt{\sum(x_i-\bar{x})^2}=48.784,(x_i-\bar{x})^2=10.726$$

$$\mu_x\in[3.063,3.340]$$

Which means the probability of the mean freshman GPA for students whose ACT score is 28 in the range of 3.063 to 3.34 is 95%.

(b) The 95% prediction interval of μ_x is

$$\mu_x\in\left[\widehat{\mu_x}-1.96\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}},\widehat{\mu_x}+1.96\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}}\right]$$

$$\widehat{\mu_x}=\widehat{\beta_0}+\widehat{\beta_1}x=2.1147+0.0388\times28=3.2011,\mathrm{n}=120$$

$$\hat{\sigma}=0.623,\sqrt{\sum(x_i-\bar{x})^2}=48.784,(x_i-\bar{x})^2=10.726$$

$$\mu_x\in[1.972,4.430]$$

Which means the probability of Mary Jones's GPA in the range of 1.972 to 4.43 is 95%.

(c) Yes, the prediction interval wider than the confidence interval. The confidence interval is an inference on a parameter, therefore, it is intended to cover the value of the parameter. The prediction interval describes the value for a random variable and therefore must have a wider interval to allow for non-parameterized variables to impact the predicted value.

(d) The 95% confidence band is

$$\widehat{\mu_x}\pm\sqrt{2F(0.05;2.118)}\hat{\sigma}\sqrt{\frac{1}{120}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

So $\mu_x \in [3.026, 3.376]$

It is wider than the interval in part (a) because it is representing the confidence intervals for the entire regression line, not just at a single point X_h.

2.51

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$E(b_0) = E(\bar{Y} - b_1\bar{X}) = \bar{Y} - \bar{X}E(b_1) = \bar{Y} - \bar{X}\beta_1 = \beta_0$$

2.52

$$\mathrm{Var}(\widehat{\beta_0}) = \mathrm{Var}(\bar{Y} - \widehat{\beta_1}\bar{X}) = \mathrm{Var}(\bar{Y}) + \mathrm{Var}(\widehat{\beta_1}\bar{X}) + 2\mathrm{Cov}(\bar{Y}, \widehat{\beta_1}\bar{X})$$

Based on (2.31)

$$\mathrm{Cov}(\bar{Y}, \widehat{\beta_1}) = 0, 2\mathrm{Cov}(\bar{Y}, \widehat{\beta_1}\bar{X}) = 2\bar{X}Cov(\bar{Y}, \widehat{\beta_1}) = 0$$

Since $Y_i, i = 1, \dots, n$ are $i.i.d$

$$Var(\bar{Y}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Y_i = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$Var(\widehat{\beta_1}\bar{X}) = \bar{X}^2 Var(\widehat{\beta_1}) = \bar{X}^2 \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$Var(\widehat{\beta_0}) = \frac{\sigma^2}{n} + \bar{X}^2 \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$Var(\widehat{\beta_0})$ is a special case when $X_h = 0$ in $Var(Y_h) = Var(\bar{Y} + \widehat{\beta_1}(X_h - \bar{X}))$

2.63

We use Python to solve this problem.

Since we don't know σ^2, the confidence interval is

$$\beta_1 \in [\widehat{\beta_1} - \mathrm{t}_{0.95}(\mathrm{n}-2)\frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}, \widehat{\beta_1} - \mathrm{t}_{0.05}(\mathrm{n}-2)\frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}]$$

You can see the code and print result below.

```python
# classify with geographic region
def cdi_ci(data, x, y):
    '''x: string; y: string; data: dataframe'''
    # linear regression model
    lm0 = LinearRegression()
    lm0.fit(X = data[x].values.reshape(-1,1), y = data[y])
    # slope
    slope = lm0.coef_[0]
    # intercept
    intercept = lm0.intercept_
    # residual
    residual = lm0.predict(data[x].values.reshape(-1,1)) - data[y].values
    # sigma
    sigma = np.sqrt(sum(residual*residual)/(len(data)-2))
    # sqrt(sum(x_i - x_bar)^2)
    s_error = np.sqrt(sum((data[x] - np.mean(data[x]))**2))
    # confidence interval
    blLow = slope - t.ppf(0.95, df = len(data)-2)*sigma/s_error
    blHigh = slope - t.ppf(0.05, df = len(data)-2)*sigma/s_error
    return blLow, blHigh
```

```python
for i in range(1, 5):
    print('confidence interval when geographic region equals to {0}'.format(i))
    print(cdi_ci(CDI[CDI['geographic_region'] == i], 'precent_bachelor_deg', 'per_capita_income'))
```

```
confidence interval when geographic region equals to 1
(460.517703881817, 583.799961242046)
confidence interval when geographic region equals to 2
(193.48578916011115, 283.853007843554)
confidence interval when geographic region equals to 3
(285.707618101109, 375.5158331337243)
confidence interval when geographic region equals to 4
(364.75848918029646, 515.8729252228775)
```