MATH3424 HW1
Name: Leung Ko Tsun
SID: 20516287

Q1a). Response variable: gasoline consumption of cars
Predictors: Number of cylinders
Explanation: Number of cylinders can be used to predict the gasoline consumption. And gasoline consumption is of primary importance as we are more interested at this number than the number of cylinders.

Q1b). Response variable: college admission
Predictors: SAT scores, grade point average
Explanation: The primary importance is the college admission figure rather than other two variables. Also, it is more intuitive to make prediction on college admission rate based on SAT scores and grade point average.

Q1c). Both supply and demand can be regarded as response variables or predictors. It depends on the usage and objective of such prediction. For example, if I am a major producer of a specific good, let say rice, then it makes sense for me to predict the demand in different production level of rice since I can control the level of supply. If I am an economist and want to predict the supply level based on different supply level, then supply should be treated as response variable and demand is the predictor.

Q1d). Response variable: Return on stock
Predictors: Company's asset, net sales
Explanation: Return on stock means the capital gain on investing the stock of a specific company. It is intuitive to make prediction on the return on stock based on the company's assets and net sales, as these predictors can show if the company is doing well on business development and asset allocation. Return on stock is of primary importance rather than the other two variables.

Q1e). Response variable: The time to run the race
Predictors: The distance of the race, weather conditions.
Explanation: The time to run the race is of primary importance than the distance and weather conditions. It is also intuitive to make predictions on the time based on distance and weather to estimate the time cost for a running competition.

Q1f). Response variable: Whether or not the person has a lung cancer
Predictors: The weight of a person, whether or not the person is a smoker
Explanation: It is intuitive to make prediction on the probability of a person having lung cancer based on the weight and smoker, as if the person is obese and smoke frequently, the probability having lung cancer increases. Whether or not the person has a lung cancer is of primary importance than smoker status or weight of that person.

Q1g). Response variable: The height and weight of a child
Predictors: Parents' height and weight, the gender and age of the child

Explanation: The height and weight of a child is of primary importance. It is intuitive to predict the height and weight of a child based on parents' height and weight, also the gender and age of the child, as to find interesting genetic relationship of height and weight.

Problem 2:

$$\bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i$$

$$\bar{x} = 1/20 * (240 + 243 + 250 + 254 + 264 + 279 + 284 + 285 + 290 + 302 + 310 + 312 + 315 + 322 + 337 + 348 + 386 + 520)$$

$$\bar{x} = 311.15$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s^2 = \frac{1}{19}\sum_{i=1}^{n}(x_i - 311.15)^2$$

$$s^2 = 4146.45$$

Hypothesis test: $H_0: \mu = 200$, $H_1: \mu \neq 200$

$$t\text{-statistics} = t_1 = \frac{\bar{x} - 200}{s/\sqrt{n}}$$

$$= \frac{311.15 - 200}{\sqrt{4146.45/20}} = 7.71945$$

From t-distribution table, $t_{(n-1,\frac{\alpha}{2})} = t_{(19,0.01)} = 2.539$

$\therefore |t_1| = 7.71945 > t_{(19,0.01)}$

So. at a significant level of 0.02, we reject H_0, μ is statistically different from 200.

R code:

```
>
> #q2
> q2data <- read.table("~/Downloads/3424/q2.txt", head = TRUE)
>
> t.test(q2data$observation, y = NULL,
+        alternative = "two.sided",
+        mu = 200,
+        paired = FALSE,
+        var.equal = FALSE,
+        conf.level = 0.98
+ )

        One Sample t-test

data:  q2data$observation
t = 7.7194, df = 19, p-value = 2.836e-07
alternative hypothesis: true mean is not equal to 200
98 percent confidence interval:
 274.5847 347.7153
sample estimates:
mean of x
   311.15
```

Problem 3

a. Disagree, since $-1 \leq Corr(Y,X) \leq 1$

b. Disagree, '.' $Cov(Y,X)=0$ or $Corr(Y,X)=0$ only shows X and Y are not linearly related. as the correlation coefficient measures only linear relationship. $Cov(Y,X)$ can be 0 even X and Y has non-linear relationship.

c. Agree. The least square regression line is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

As showed in lectures, the least square estimates

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

$$\Rightarrow \quad \hat{Y} = \bar{y} + \hat{\beta}_1 (X - \bar{x})$$

When $x = \bar{x}$, $\hat{Y} = \bar{y} + \hat{\beta}_1 (\bar{x} - \bar{x})$

$$\hat{Y} = \bar{y}$$

So the least square regression line passes through $(\bar{x}, \bar{y})$.

3d). Agree. Let $\hat{\beta}_{0(Y,\hat{Y})}$ and $\hat{\beta}_{1(Y,\hat{Y})}$ be the intercept and slope of the least square line in the graph of Y versus $\hat{Y}$

$$\hat{\beta}_{1(Y,\hat{Y})} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}$$

$$= \frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{y})}{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2} \quad (\bar{y}=\bar{\hat{y}}, \text{ by (c)})$$

$$= \frac{\sum_{i=1}^{n}[(\hat{y}_i-\bar{y})+(y_i-\hat{y}_i)](\hat{y}_i-\bar{y})}{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}$$

$$= \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})}{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}$$

$$= \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}\hat{y}_i(y_i-\hat{y}_i)-\bar{y}\sum_{i=1}^{n}(y_i-\hat{y}_i)}{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}$$

$$= \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}$$

($\sum_{i=1}^{n}\hat{y}_i(y_i-\hat{y}_i)=0 \because \hat{Y}\perp(\hat{Y}-Y)$, proved in lecture.
$\bar{y}\sum_{i=1}^{n}(y_i-\hat{y}_i)=0 \because$ sum of residual $=0$)

$$= 1$$

$$\hat{\beta}_0(\bar{y},\bar{y}) = \bar{y} - \hat{\beta}_1 \bar{y}$$
$$= \bar{y} - 1 \times \bar{y}$$
$$= 0$$

The following questions are typed in latex as I don't have pen or electronic pen at the moment when I am doing the homework

1 Problem 4(a)

a.

$$H_0 : \beta_1 = 15 \; versus \; \beta_1 \neq 15$$

The t-statistics

$$t_1 = \frac{\hat{\beta}_1 - 15}{se(\hat{\beta}_1)}$$
$$t_1 = \frac{15.509 - 15}{0.505}$$
$$t_1 = 1.00792$$

From the distribution table, $t_{(n-2,\frac{\alpha}{2})} = t_{(12,0.025)} = 2.179$
$\therefore |t_1| = 1.00792 < t_{(12,0.025)}$
Therefore, at a significant level of 0.05, we do not reject the null hypothesis, i.e. β_1 is not statistically different from 15.

2 Problem 4(b)

b.

$$H_0 : \beta_1 = 15 \; versus \; \beta_1 > 15$$

The t-statistics

$$t_1 = \frac{\hat{\beta}_1 - 15}{se(\hat{\beta}_1)}$$
$$t_1 = \frac{15.509 - 15}{0.505}$$
$$t_1 = 1.00792$$

From the distribution table, $t_{(n-2,\alpha)} = t_{(12,0.05)} = 1.782$
$\therefore |t_1| = 1.00792 < t_{(12,0.005)}$
Therefore, at a significant level of 0.05, we do not reject the null hypothesis, i.e. β_1 is not statistically larger from 15.

3 Problem 4(c)

c.

$$H_0 : \beta_0 = 0 \; versus \; \beta_0 \neq 0$$

The t-statistics

$$t_1 = \frac{\hat{\beta_0}}{se(\hat{\beta_0})}$$

$$t_1 = \frac{4.162}{3.355}$$

$$t_1 = 1.24054$$

From the distribution table, $t_{(n-2,\frac{\alpha}{2})} = t_{(12,0.025)} = 2.179$
$\therefore |t_1| = 1.24054 < t_{(12,0.025)}$
Therefore, at a significant level of 0.05, we do not reject the null hypothesis, i.e. β_0 is not statistically different from 0.

4 Problem 4(d)

d.

$$H_0 : \beta_0 = 5 \; versus \; \beta_0 \neq 5$$

The t-statistics

$$t_1 = \frac{\hat{\beta_0} - 5}{se(\hat{\beta_0})}$$

$$t_1 = \frac{4.162 - 5}{3.355}$$

$$t_1 = -0.24978$$

From the distribution table, $t_{(n-2,\frac{\alpha}{2})} = t_{(12,0.025)} = 2.179$
$\therefore |t_1| = 0.24978 < t_{(12,0.025)}$
Therefore, at a significant level of 0.05, we do not reject the null hypothesis, i.e. β_0 is not statistically different from 5.

5 Problem 5

The 98% confidence interval for β_0

$$= [\hat{\beta}_0 – t_{(n-2,\frac{\alpha}{2})} \times se(\hat{\beta}_0), \hat{\beta}_0 + t_{(n-2,\frac{\alpha}{2})} \times se(\hat{\beta}_0)]$$

$$= [4.162 – t_{(12,0.01)} \times 3.355, 4.162 + t_{(12,0.01)} \times 3.355]$$

$$= [4.162 – 2.681(3.355), 4.162 + 2.681(3.355)]$$

$$= [-4.832755, 13.156755]$$

6 Problem 6

a. By definition, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
$\hat{\beta}_0 = \bar{y} – 0$
$\hat{\beta}_0 = \bar{y}$

b. The ordinary least square residuals

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

$$= \sum_{i=1}^{n} (y_i - \beta_0)$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})$$

, which is the sum of deviation from mean

c. The ordinary least square residuals

$$= \sum_{i=1}^{n} (y_i - \bar{y})$$

$$= \sum_{i=1}^{n} y_i - n\bar{y}$$

$$= \sum_{i=1}^{n} y_i - n \times \frac{1}{n} \sum_{i=1}^{n} y_i (by\ definition\ of\ \bar{y})$$

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i$$

$$= 0$$

7 Problem 7

a.

$$\bar{y} = \frac{1}{14}(23+29+49+64+74+87+96+97+109+119+149+145+154+166)$$

$$\bar{y} = 97.21$$

$$\bar{x} = \frac{1}{14}(1+2+3+4+4+5+6+6+7+8+9+9+10+10)$$

$$\bar{x} = 6$$

$$Cor(Y,X)$$

$$= \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$= \frac{1768}{\sqrt{27768.34\times 114}}$$

$$\approx 0.993698695$$

$$\approx 0.994$$

$$Cor(Y,\hat{Y})$$

$$= \frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}}$$

$$= \frac{27419.1071}{\sqrt{27768.3571\times 27418.5783}}$$

$$\approx 0.993698$$

$$\approx 0.994$$

b. SST
$= \sum_{i=1}^{n}(y_i-\bar{y})^2$
$= 5507.76+4653.19+2324.62+1103.19+538.9+104.33+1.47+0.05+138.9+474.62+2681.76+2283.47+3224.62+4731.47$
$= 27768.36$

c. SSE
$= \sum_{i=1}^{n}(y_i-\hat{y})^2$
$= 3.33^2+(-6.18)^2+(-1.69)^2+(-2.2)^2+7.8^2+5.29^2+(-1.21)^2+(-0.21)^2+(-3.72)^2+(-0.23)^2+5.26^2+1.26^2+(-5.25)^2+6.75^2$
$= 348.848$

8 Problem 8

a. Var(Y)

$$= \frac{1}{n-1} \sum_{(} i = 1)^{n} (y_i - \bar{y})^2$$

$$= \frac{1}{n-1} SST$$

$$= \frac{1}{n-1} (SSR + SSE)$$

$$= \frac{1}{19-1} (0.0358 + 0.0544)$$

$$= \frac{451}{90000}$$

$$\approx 0.0050111$$

As shown in lectures, $[Cor(Y,X)]^2 = R^2$

$$Cor(Y,X) = \sqrt{R^2}$$
$$Cor(Y,X) = \sqrt{0.397}$$
$$\approx 0.63008$$

b. The estimated participation rate, $\hat{y_0}$

$$= \hat{\beta}_0 + \hat{\beta}_1 x_0$$
$$= 0.203311 + 0.65604(0.45)$$
$$= 0.498529$$
$$= 49.8529\%$$

c. The standard error of the prediction, $se(\hat{y_0})$

$$= \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$
$$= \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)Var(X)}}$$
$$= (0.0566)\sqrt{\frac{1}{19} + \frac{(0.45 - 0.5)^2}{(19-1)(0.005)}}$$
$$= 0.0160498037$$

The 95% confidence interval of $\hat{y_0}$

$$= [\hat{y_0} - t_{(n-1,\frac{\alpha}{2})} se(\hat{y_0}), \hat{y_0} + t_{(n-1,\frac{\alpha}{2})} se(\hat{y_0})]$$

$$= [\hat{y_0} - t_{(17,0.025)} se(\hat{y_0}), \hat{y_0} + t_{(17,0.025)} se(\hat{y_0})]$$

$$= [0.498529 - 2.11(0.0160498037), 0.498529 + 2.11(0.0160498037)]$$

$$= [0.464663914191, 0.532394085808]$$

$$\approx [46.466\%, 53.239\%]$$

d. The 95% confidence interval for β_1

$$= [\hat{\beta_1} - t_{(n-2,\frac{\alpha}{2})} se(\hat{\beta_1}), \hat{\beta_1} + t_{(n-2,\frac{\alpha}{2})} se(\hat{\beta_1})]$$

$$= [0.65604 - 2.11(0.1961), 0.65604 = 2.11(0.1961)]$$

$$= [0.242269, 1.069811]$$

e. $H_0 : \beta_1 = 1$ *versus* $H_1 : \beta_1 > 1$

The t-statistic, $t_1 = \frac{\hat{\beta_1} - 1}{se(\hat{\beta_1})}$

$$= \frac{0.65604 - 1}{0.1961}$$

$$= -1.75400306$$

From the t-distribution table, $t_{(n-2,\alpha)} = t_{(17,0.02)} = 2.224$

$\therefore t_1 = -1.75400306 < t_{(17,0.02)}$

Therefore, at a significant level of 0.02, we do not reject the null hypothesis, i.e. β_1 is not statistically larger than 1.

f. Since $R^2 = [Cov(Y, X)]^2$, if Y and X were reversed in the above regression, covariance or correlation will not be affected, so R^2 is expected to remain the same.

9 Problem 9

a. Example: Y = distance travelled of a car, X = time elapsed from the start of driving. When x = 0, distance travelled must be 0.

b. Least squares estimate of β_1

$$min_{\beta_1} \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$$

Taking derivative with respect to β_1,

$$2\sum_{i=1}^{n} (y_i - \beta_1 x_i)(-x_i) = 0$$

$$-\sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n} \beta_1 x_i^2 = 0$$

$$\beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

$$\beta_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

c.In simple linear regression with an intercept, we obtain β_0 and β_1 by doing the minimization

$$min_{\beta_0,\beta_1} \sum_{i=1}^{n} [y_i - (\beta_1 x_i + \beta_0)]^2$$

And taking the gradient with respect to β_0 and β_1, solving for β_0 and β_1.

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} [y_i - (\beta_1 x_i + \beta_0)]^2 = -2\sum_{i=1}^{n} [y_i - (\beta_1 x_i + \beta_0)] = 0$$

$$\sum_{i=1}^{n} [y_i - (\beta_1 x_i + \beta_0)] = 0$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

$$\sum_{i=1}^{n} e_i = 0$$

However, in the no-intercept model setting, as there is no constant term giving the above condition, the residuals will not necessarily sum up to 0.

Problem 10 (R is used to obtain the solutions)

a. Cov(husband, wife) = 69.41294 cm^2
b. Cov(husband, wife) = 10.75903 in^2
c. cor(husband, wife) = 0.7633864
d. cor(husband, husband – 5) = 1
e. wife = 41.93015 + 0.69965 * husband
f. p-value.= 0.000000000000000000001536359 < 0.05
   At the significant level of 0.05, the null hypothesis is rejected, i.e., the slope is statistically different from 0.
g. p-value = 0.00016 < 0.05
   At the significant level of 0.05, the null hypothesis is rejected, i.e., the intercept is statistically different from 0.

```
> #q10
> heights <- read.table("~/Downloads/3424/q10.txt", header=TRUE)
> husband <- heights$Husband
> wife <- heights$Wife
> cov(husband, wife)
[1] 69.41294
> cov(husband * 0.393700787, wife * 0.393700787)
[1] 10.75903
> cor(husband, wife)
[1] 0.7633864
> cor(husband, husband - 5)
[1] 1
> regression <- lm(wife ~ husband, data = heights)
> summary(regression)

Call:
lm(formula = wife ~ husband, data = heights)

Residuals:
     Min       1Q   Median       3Q      Max
-19.4685  -3.9208   0.8301   3.9538  11.1287

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.93015   10.66162   3.933 0.000161 ***
husband      0.69965    0.06106  11.458  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.928 on 94 degrees of freedom
Multiple R-squared:  0.5828,    Adjusted R-squared:  0.5783
F-statistic: 131.3 on 1 and 94 DF,  p-value: < 2.2e-16

> summary(regression)$coefficients[1:2,]
              Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 41.9301535 10.66162266  3.932812 1.605824e-04
husband      0.6996537  0.06106163 11.458156 1.536359e-19
```