ESI 4606 Analytics I - Foundations of Data Science
Homework 3
Due: September 28st (11:00AM), 2022

Problem 1 (2.5 points)

Consider the following data on the propagation velocity of an ultrasonic stress wave through a substance, y (km/s), and the tensile strength of substance, x (MPa).

Table 1: Hypothetical data on the propagation velocity

x, MPa	12	30	36	40	45	57	62	67	71	78	93	94	100	105
y, km/s	3.3	3.2	3.4	3.0	2.8	2.9	2.7	2.6	2.5	2.6	2.2	2.0	2.3	2.1

Suppose a simple linear regression, i.e., $y = \beta_0 + \beta_1 x + \varepsilon$, is used to fit the data, where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. Least squares estimation is employed to estimate the model parameters. Compute the following through **hand calculation**.

(a) What are estimated values for $\hat{\beta}_0$ and $\hat{\beta}_1$? What are their corresponding interpretations.

Solution: Simplify the estimate:

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y})}{\sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2)} \\
&= \frac{\sum x_i y_i - \bar{y}\sum x_i - \bar{x}\sum y_i + n\bar{x}\bar{y}}{\sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2} \\
&= \frac{\sum x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}}{\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2} \\
&= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{2234.3 - 14 * 63.57 * 2.69}{67182 - 14 * 4041.33} \\
&= -0.01471 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
&= 2.69 - (-0.01471) * 63.57 = 3.62091
\end{aligned}
$$

$\hat{\beta}_1$ is the slope of the fitted line and $\hat{\beta}_0$ is the intercept.

(b) For a two-sided hypothesis test: $\mathrm{H}_0 : \beta_1 = 0$ v.s. $\mathrm{H}_1 : \beta_1 \neq 0$, use t-test approach and a significance level of $\alpha = 0.05$ to perform the hypothesis testing and draw the conclusion.

Solution:

$\hat{\sigma}^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{0.26246}{14-2} = 0.02187$

$SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}} = \sqrt{\frac{0.02187}{10603.43}} = 0.001436$

$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta})_1} = \frac{-0.01471}{0.001436} = -10.2429$

critical-t value $\cancel{t_{.975,12}}$: $t_{0.025,12}$

```
> qt(.975, 12)
[1] 2.178813
```

Because $|t| > t_{.975,\ 12}$, reject H_0 based on significant level 0.05.

(c) What is the 95% confidence interval for β_1? What is the corresponding interpretation?

Solution:

$$
\begin{aligned}
CI &= [\hat{\beta}_1 - \underbrace{\cancel{t_{1-\alpha/2,n-2}}}_{t_{\alpha/2,n-2}} * SE(\hat{\beta}_1), \hat{\beta}_1 + \underbrace{\cancel{t_{1-\alpha/2,n-2}}}_{t_{\alpha/2,n-2}} * SE(\hat{\beta}_1)] \\
&= [-0.01471 - 2.1788 * 0.001436, -0.01471 + 2.1788 * 0.001436] \\
&= [-0.01784, -0.01158]
\end{aligned}
$$

Parameter $\hat{\beta}_0$ is likely to reside in this interval with a confidence level 95%.

(d) What are values for R^2 and $\hat{\sigma}$? If tensile strength of substance is 50 MPa, what is the predicted propagation velocity of an ultrasonic stress wave through the substance based on this model?

Solution:

$\hat{\sigma} = \sqrt{0.02187} = 0.14789$

$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{0.26246}{2.55714} = 0.89736$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 3.62091 - 0.01471 * 50 = 2.885$

Problem 2 (1 point)

Prove that the fitted least squares line, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, will always pass through the point ($\bar{x}$,$\bar{y}$), where $\bar{x}$ and $\bar{y}$ are sample averages.

Proof:

Recall $\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

When $x = \bar{x}$, RHS $= \hat{\beta_0} + \hat{\beta}_1 \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y} =$ LHS. Thus, the regression line will always pass through $(\bar{x}, \bar{y})$.

Problem 3 (1.5 points)

This question involves **using R** to perform the multiple linear regression using the "**Carseats**" datain from R library of "**ISLR**".

(a) Fit a multiple regression model to predict "Sales" using "Price", "Urban", and "US". Use the summary() function to print the results.

Solution:

```
> library(ISLR2)
> attach(Carseats)
> model_a <- lm(Sales~Price+Urban+US)
> summary(model_a)

Call:
lm(formula = Sales ~ Price + Urban + US)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9206 -1.6220 -0.0564  1.5786  7.0581

Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
Price       -0.054459   0.005242 -10.389  < 2e-16 ***
UrbanYes    -0.021916   0.271650  -0.081    0.936
USYes        1.200573   0.259042   4.635 4.86e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2335
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b) Provide an interpretation of each coefficient in the model. It is noted that some of the input variables in the model are qualitative.

Interpretation:

Coefficients are in the **Estimate** column in the model summary.

The coefficient of the intercept represents the mean value of the response "Sales" when all of the predictor variables in the model are equal to 0.

Coef. of variable "Price" means the expected change in the outcome "Sales" with increasing "Price" by 1 unit.

Coef. of variable "UrbanYes" means the expected change in the outcome "Sales" with flipping the variable "Urban" from "No" to "Yes", similar for "USYes".

(c) Based on significant level of $\alpha = 0.05$, for which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

Solution: (This solution is still based on $model_a$)

Price and USYes. Because they have P-values < 0.05, in this case, we **reject** the H0.

(d) On the basis of your response to question (c), fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

Solution:

```
> model_d <- lm(Sales~Price+US)
> summary(model_d)

Call:
```

```
lm(formula = Sales ~ Price + US)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9269 -1.6286 -0.0574  1.5766  7.0515 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
Price       -0.05448    0.00523 -10.416  < 2e-16 ***
USYes        1.19964    0.25846   4.641 4.71e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.469 on 397 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2354 
F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

(e) How well do the models in (a) and (d) fit the data?
Solution:
Firstly, both of the models are statistically significant according to the F-test result. $R^2 = 0.2393$ for $model_a$ and $model_d$, means about 23.93% of the variability observed in the response can be explained by either of the two models.
According to the RSE values, for $model_a$, it is 2.472 while for $model_d$ is 2.469, implying they fit the data almost equally.
Considering the R^2's are small and RSE values are not very small, neither of them fits the data very well.

(f) Using the model from (d), obtain 95% confidence intervals for the coefficient(s).

```
> confint(model_d)
                  2.5 %      97.5 %
(Intercept) 11.79032020 14.27126531
Price       -0.06475984 -0.04419543
USYes        0.69151957  1.70776632
```

Note: **To get full points, include R codes in the appendix sections**