Business Statistics Midterm Exam

Fall 2019: BUS41000

This is a closed-book, closed-notes exam. You may use any calculator.

Please answer all problems in the space provided on the exam.

Read each question carefully and clearly present your answers.

Honor Code Pledge: "I pledge my honor that I have not violated the University Honor Code during this examination."

Sign: _____

Name: _____

Useful formulas

- $E(aX + bY) = aE(X) + bE(Y)$
- $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2ab \cdot Cov(X, Y)$
- $Cor(X, Y) = \frac{Cov(X,Y)}{sd(X) \cdot sd(Y)}$
- The standard error of $\bar{X}$ is defined as $s_{\bar{X}} = \sqrt{\frac{s_X^2}{n}}$, where s_X^2 denotes the sample variance of X.
- The standard error for the difference in the averages between groups a and b is defined as:

$$s_{(\bar{X}_a - \bar{X}_b)} = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$$

  where s_a^2 denotes the sample variance of group a and n_a the number of observations in group a.
- The standard error for a proportion is defined by: $s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- The standard error for difference in proportion is defined by:

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

  where $\hat{p}_1$ and $\hat{p}_2$ denote two independent proportions, and n_1 and n_2 are the number of trials.
- Bayes's formula:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

  where A, B are two events.
- For $Z \sim N(0, 1)$, $P(-1 \leq Z \leq 1) = 68\%, P(-2 \leq Z \leq 2) = 95\%, P(-3 \leq Z \leq 3) = 99\%$.
- Similarly, $X \sim N(\mu, \sigma^2)$, $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 95\%$.
- Standardization to standard normal: assume $X \sim N(\mu, \sigma^2)$, $Z \sim N(0, 1)$, then

$$P(a \leq X \leq b) = P(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}).$$

- The Sharpe Ratio for Stock S: $\frac{E(S)}{\sqrt{Var(S)}}$.

Grading Sheet for TA:

Problem	Score
P1	
P2	
P3	
P4	
P5	
P6	
P7	
P8	
P9	
Total	

Problem 1: B-17 flying fortress and Wald. [15 points]

During a period in World War II, the U.S. Army Air Forces (AAF) would send over 300 B-17 bombers daily to raid factories in Germany. These missions, originating in the U.K., were dangerous. In the peak of the campaign, the return probability for a B-17 crew was only 80%.

In trying to reduce the probability of a failed mission, a Navy statistician, Abraham Wald, was put in charge of studying the damage patterns in the B-17's that successfully made back from a mission. His ultimate goal was to decide where to add extra armor in the planes (you could not just add heavy armor everywhere, as the planes would be too heavy to fly!). Wald was able to learn that if a plane made back from a mission, there was a 67% probability that it was shot in the fuselage, 15% in the fuel systems, 10% in the cockpit area and 8% in the engines.

From experiments, Wald was also able to deduce that during combat, a B-17 would be shot in the fuselage with 56% probability, in the fuel systems with 14%, in the cockpit area 14% and engine 16%.

1. Based on this information, what was Wald's recommendation to the AAF, i.e., if they had to choose one area of the plane, where should they add extra armor to the B-17's? (Hint: Wald suggested to improve on the weakest area: the area with the smallest returning probability given it is shot.) [10 points]

$$\begin{aligned} P(\text{success return} \mid \text{area being shot}) &= \frac{P(\text{success return\&area being shot})}{P(\text{area being shot})} \\ &= \frac{P(\text{area being shot} \mid \text{success return})P(\text{success return})}{P(\text{area being shot})} \end{aligned}$$

So

$$\begin{aligned} \text{fuselage}: &\frac{0.67 \times 0.80}{0.56} = 0.957 \\ \text{fuel}: &\frac{0.15 \times 0.80}{0.14} = 0.857 \\ \text{cockpit}: &\frac{0.1 \times 0.80}{0.14} = 0.571 \\ \text{engines}: &\frac{0.08 \times 0.80}{0.16} = 0.40 \end{aligned}$$

So engines are most weak part, needs extra armor.

2. Can you calculate the probabilities of being hit in the fuel systems and in the engines, respectively, given the plane did not return? [5 points]

$$\begin{aligned} P(\text{area being shot} \mid \text{not return}) &= \frac{P(\text{not return\&area being shot})}{P(\text{not return})} \\ &= \frac{P(\text{not return} \mid \text{area being shot})P(\text{area being shot})}{P(\text{not return})} \end{aligned}$$

$$\begin{aligned} \text{fuel}: &\frac{(1-0.857) \times 0.14}{0.2} = 0.1001 \\ \text{engines}: &\frac{(1-0.4) \times 0.16}{0.2} = 0.48 \end{aligned}$$

Problem 2: Choosing an agent. [10 points]

You are considering to purchase a house. On a rating site, you have collected data on the two potential real estate agents in Chicago. For each rating, there are only two categories, YES (recommend) or NO (not recommend).

Recommend?	Agent BIG	Agent SMALL
YES	1644	192
NO	548	48

Is the Agent SMALL better? Justify your answer using either hypothesis testing or confidence interval (with 95% confidence guarantee). [10 points]

$$\frac{1644}{1644+548} = 0.75, \quad \frac{192}{192+48} = 0.8$$

Difference is

$$0.75 - 0.8 = -0.05$$

Standard deviation is

$$s = \sqrt{\frac{0.75(1-0.75)}{1644+548} + \frac{0.8(1-0.8)}{192+48}} = 0.0274$$

Confidence interval at 95% level

$$[-0.05 \pm 2 \times 0.0274] = [-0.1048, 0.0048]$$

The confidence interval contains 0, we cannot reject the null hypothesis.

Or by hypothesis testing

$$\frac{0.05}{0.0274} = 1.825 < 2$$

We cannot reject the null hypothesis.

Problem 3: Which insurance to purchase? [10 points]

The next step is to choose a house insurance policy. Suppose there are three options available: standard policy, premium policy, and no policy (not insured). If you decide on a policy, you will have to buy it for the whole year.

Policy	Cost per month	Deductible if you file claim for house damage
Standard	\$50	\$5000
Premium	\$55	\$500

Suppose in one year, there is a 1% chance of house damage, and you estimate that the damage will cost you \$200,000.

1. For one year, which one of the three options you would like to choose in expectation? Which option has the smallest amount of variability? [5 points]

If the house is damaged, you pay deductible if you have insurance, otherwise 200,000.

Standard

$$50 * 12 + 5000 * 0.01 = 650$$

Premium

$$55 * 12 + 500 * 0.01 = 665$$

No insurance

$$200000 * 0.01 = 2000$$

Variance

Standard

$$0.99 * (600 - 650)^2 + 0.01 * (5600 - 650)^2 = 247500$$

Premium

$$0.99 * (660 - 665)^2 + 0.01 * (1160 - 665)^2 = 2475$$

No insurance

$$0.99 * (0 - 2000)^2 + 0.01 * (200000 - 2000)^2 = 3.96 * 10^8$$

2. Now suppose you want to stick to a policy for two years. The insurance company is currently running a promotion: if you do not file a claim in the first year, your monthly cost will be zero; otherwise, your monthly fee will stay the same. Suppose the probability of house damage is 1% each year and is independent. Now, which policy you prefer in expectation? [5 points]

Four possible cases: no damage happens, one damage happens in the first year, one damage happens in the second year and damages happen in both years, with probability 0.9801, 0.0099, 0.0099 and 0.0001 respectively.

Standard

$$0.9801 * 50 * 12 + 0.0099 * (50 * 12 + 5000) + 0.0099 * (50 * 24 + 5000) + 0.0001 * (50 * 24 + 5000 * 2) = 706$$

Premium

$$0.9801 * 55 * 12 + 0.0099 * (55 * 12 + 500) + 0.0099 * (55 * 24 + 500) + 0.0001 * (55 * 24 + 500 * 2) = 676.6$$

No insurance

$$0.01 * 200000 * 2 = 4000$$

Problem 4: Portfolio. [10 points]

I am building a portfolio composed of SP500 and Bonds. Assume that $SP500 \sim N(11, 19^2)$ and $Bonds \sim N(4, 6^2)$. Here we measure the annual return in percentage (i.e., the Bond has an expected annual return of 4%, with a standard deviation of 6%).

1. Consider the 50-50 split between SP500 and Bonds, assume the standard deviation of this 50-50 portfolio is

$$sd(0.5SP500 + 0.5Bonds) = 11.000$$

Can you figure out the covariance between SP500 and Bonds, as well as the correlation? [3 points] Because

$$Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2 \times a \times b \times Cov(X, Y)$$

So

$$11^2 = 0.5^2 \times 19^2 + 0.5^2 \times 6^2 + 2 \times 0.5 \times 0.5 \times Cov(X, Y)$$

Solve the equation above, we have $Cov(X, Y) = 43.5$

$$Corr(X, Y) = \frac{Cov(X, Y)}{SD(X) \times SD(Y)} = \frac{43.5}{19 \times 6} = 0.382$$

2. Using the covariance you calculated in sub-problem 1, can you calculate

$$sd(0.8SP500 + 0.2Bonds) = ?$$

Also, which portfolio is better: the 80-20 split between SP500 and Bonds, or the 50-50 split? Justify your answer. [2 points] The variance of 80/20 portfolio is

$$0.8^2 \times 19^2 + 0.2^2 \times 6^2 + 2 \times 0.8 \times 0.2 \times (43.5) = 246.4 \approx 15.70^2$$

Sharp ratio of 50/50 portfolio is $\frac{7.5}{11} = 0.68$ and 80/20 portfolio is $\frac{9.6}{15.70} = 0.61$. 50/50 portfolio has larger Sharp ratio so we prefer it.

3. Suppose that you decide to invest \$50,000 in a 50-50 split portfolio based on SP500 and Bonds, at the beginning of 2020. By the end of 2020, you would need to pay for the property tax, which follows a normal distribution with a mean \$9,750, and a standard deviation \$2,398. What is the probability that the return of your portfolio would be enough to cover your 2020's property tax? [5 points] My return

will be $X = 50000 * (r/100) = 500r$ where $r \sim N(7.5, 11^2)$. So $X \sim N(3750, 5500^2)$. The tax follows $T \sim N(9750, 2398^2)$. Suppose X and T are independent,

$$X - T \sim N(-6000, 5500^2 + 2398^2)$$

$$P(X - T > 0) = P(Z > \frac{6000}{\sqrt{5500^2 + 2398^2}}) \approx P(Z > 1) = 0.16$$

Problem 5: Confidence interval and hypothesis testing. [15 points]

The following table summarizes the annual returns on the SP500 from 1900 until the end of 2015, in total of 116 years (in percentage terms):

116 years of SP500	
Sample average	7.2
Sample std. deviation	13.0

1. Based on these results, what is the probability of the SP500 returning less than 20% next year? In addition, give a 95% prediction interval for next year's SP500 return. [3 points]

$$P(X < 0.2) = P\left(Z < \frac{0.2 - 0.072}{0.13}\right) = P(Z < 0.98) \approx 0.84$$

The predictive interval is

$$[7.2 - 2 * 13, 7.2 + 2 * 13] = [-18.8, 33.2]$$

2. Use a 99% confidence interval, to test the hypothesis that the expected return (true mean) of the SP500 is equal to 4% a year. [2 points]

The standard deviation is $\sqrt{13^2/116} = 1.21$, confidence interval is $7.2 \pm 3 \times 1.21 = [3.58, 10.82]$. We see that 4% is inside the interval so cannot reject the null hypothesis.

3. In addition, suppose the 95% confidence interval (constructed based on our dataset) for the population mean of SP500 return μ is $[4.7, 9.6]$. Which one below best describes the statistical meaning? [5 points]

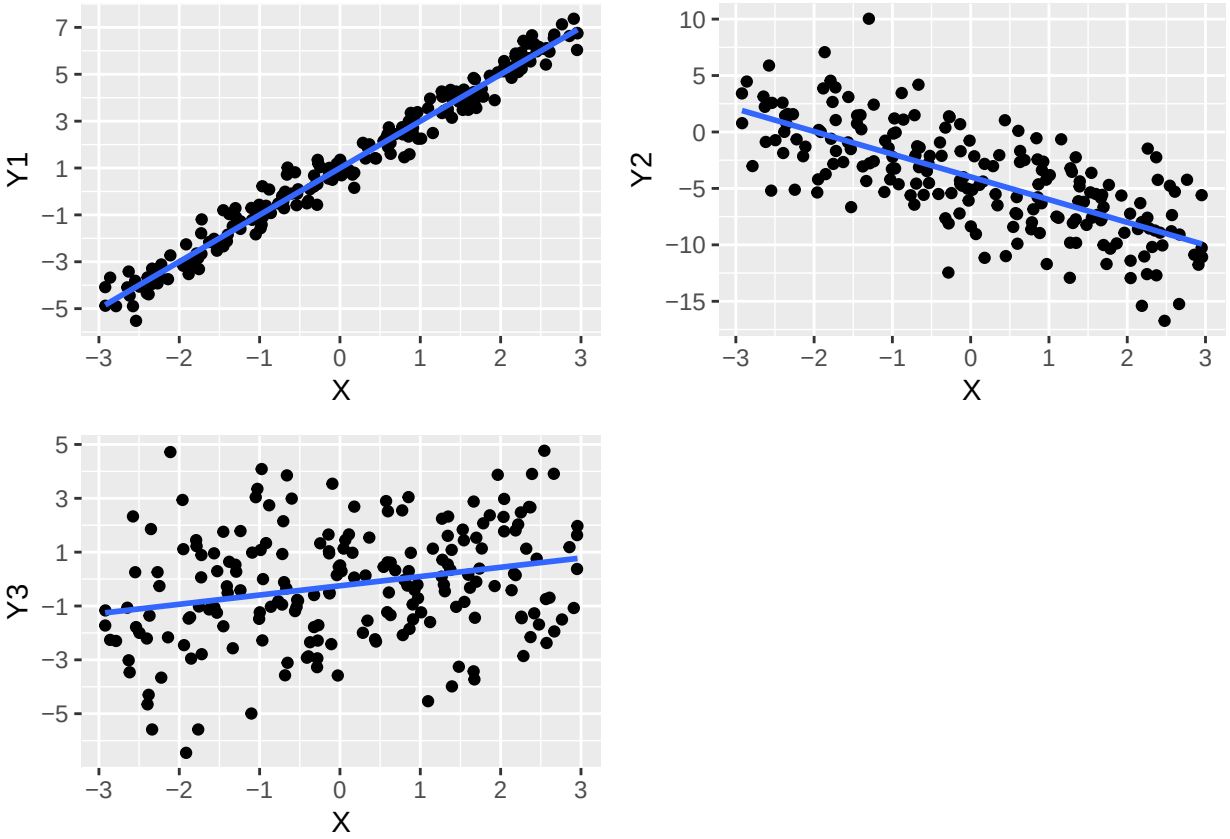
- (a) $P(\mu$ lies in $[4.7, 9.6]) = 95\%$, in other words, the probability that true mean of SP500 μ lies in the interval $[4.7, 9.6]$ is 95%.
- (b) If we recollect datasets and build confidence intervals many times, 95% of the times, these intervals will cover the true μ. **Correct**
- (c) We are 95% sure that the true mean is in the interval $[4.7, 9.6]$.

4. We want to test the null hypothesis $H_0 : \mu = 11$ vs. $H_1 : \mu \neq 11$. We calculate the t-statistics, which is $t = -3.48$. Which one below describes the statistical meaning? [5 points]

- (a) We reject the null hypothesis with 95% confidence. Here the 95% confidence means that when the null is wrong, the probability of correctly rejecting the null is 95%.
- (b) We reject the null hypothesis with 95% confidence. Here the 95% confidence means that when the null is correct, the probability of wrongfully rejecting the null is 5%. **Correct**
- (c) Both (a) and (b).

A 95% confidence level means that there is a 5% chance that your test results are the result of a type 1 error (false positive).

Problem 6: Regression. [15 points]



In the above scatterplots, three different variables $Y1, Y2, Y3$ are regressed onto the same X (in all three scatterplot we have the exact same n = 200 values for X). The line is the least square regression line. In this question, we can think of residual standard error (s) for each regression as the uncertainty of the error term, $Y = b_0 + b_1 X + \epsilon, \epsilon \sim N(0, s^2)$.

Carefully examine the plots and answer the questions below:

1. Which of the following is the least square estimates of the slope (b_1) and intercept (b_0) for the regression of $Y3$ on X? [3 points]

- (a) $b_1 = 0.34, b_0 = -0.24$ **Correct**
- (b) $b_1 = 0.78, b_0 = 0.02$
- (c) $b_1 = 2.53, b_0 = -0.05$

2. Which of the following is the least square estimates of the slope (b_1) and residual standard error (s), for regression $Y2$ on X? [2 points]

- (a) $b_1 = -0.9, s = 6.3$
- (b) $b_1 = -2.0, s = 3.2$ **Correct**
- (c) $b_1 = -4.3, s = 3.1$

3. Which of the following is the correlation (R) and residual standard error (s), for regression $Y1$ on X? [3 points]

- (a) $R = 0.988, s = 0.5$ **Correct**
- (b) $R = 0.707, s = 0.97$
- (c) $R = 0.261, s = 0.1$

4. What is the correlation between $Y2$ and X? [2 points]

- (a) -0.71 **Correct**
- (b) -0.97
- (c) -0.26

5. Using all the information provided so far, give a rough approximation for the 99% prediction interval for $Y1$ given $X = 0$. [2 points]

$$Y_1 \mid X \sim N(1, 0.5)$$
$$[1 \pm 3 \times 0.5] = [-0.5, 2.5]$$

6. What is the residual standard error s for $Y3$?
   - (a) 2.05 **Correct**
   - (b) 0.49
   - (c) 3.50

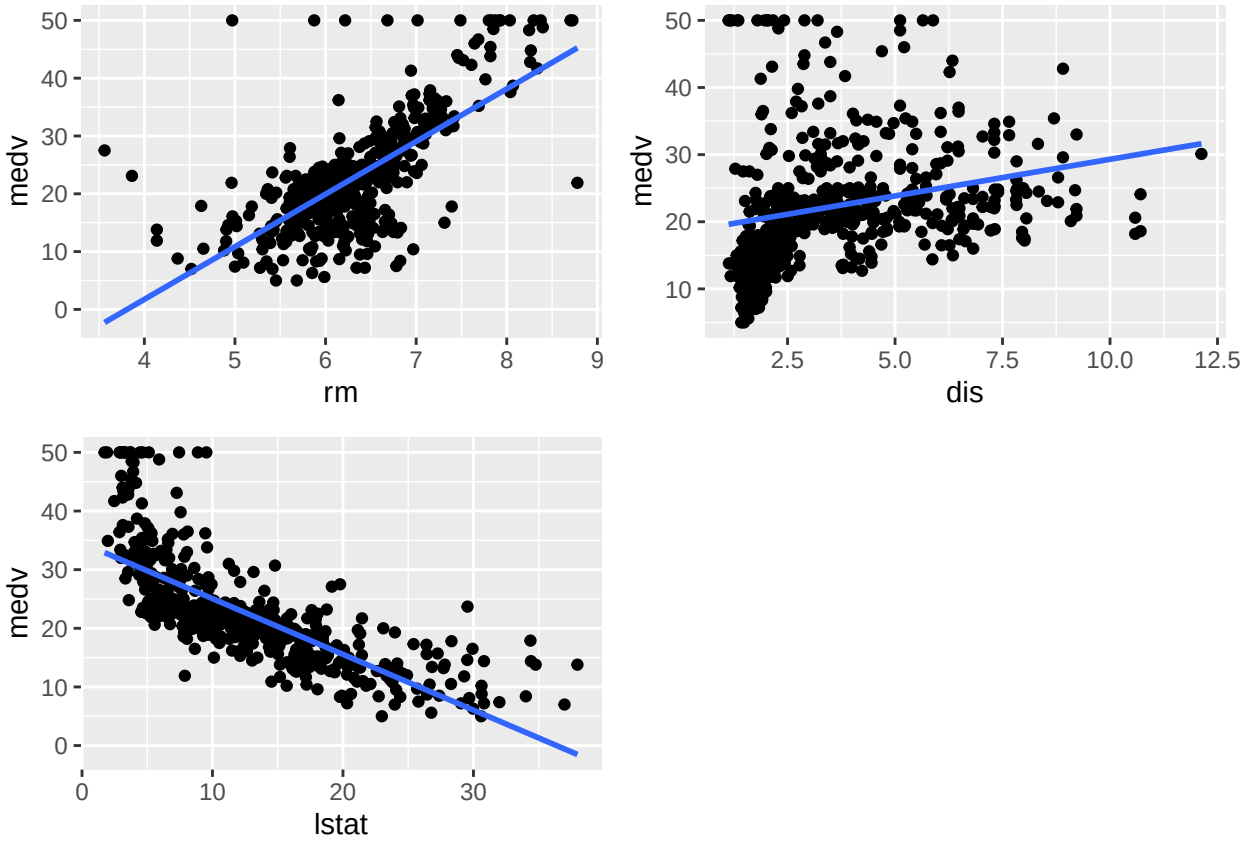
   In addition, give an approximation for $P(Y3 > 4|X = 3)$. [3 points]

Problem 7: Boston housing data. [10 points]

In this question, we take a look at a dataset that contains information collected by the U.S Census Service concerning housing in the area of Boston Massachusetts. In total, there are 506 areas in the dataset.

variables	
medv	median value of owner-occupied homes in $1000s
rm	average number of rooms per dwelling
dis	weighted mean of distances to five Boston employment centers
lstat	lower status of the population (percent)

We run three simple linear regression, aiming to figure out the quality of **rm**, **dist**, **lstat** in explaining **medv**.

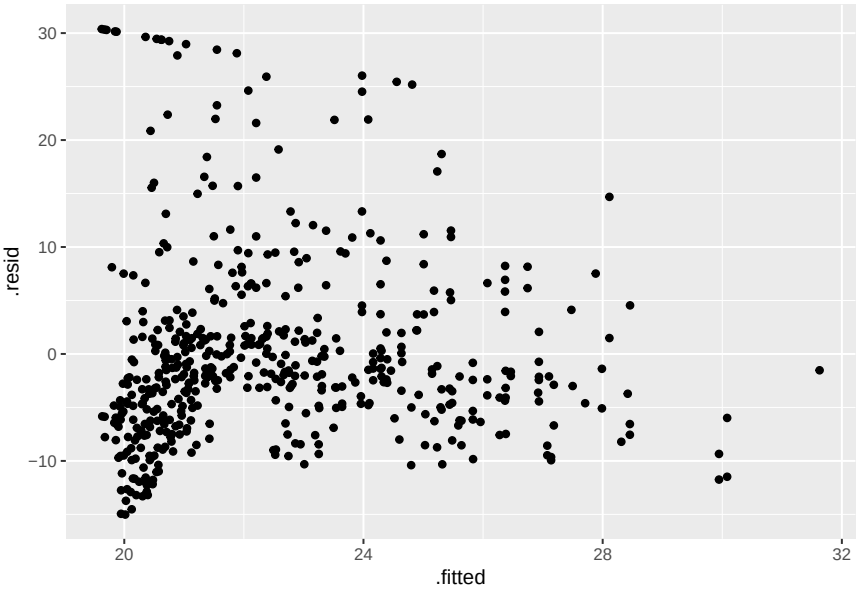


1. Based on the least square regression plot, which one of the variables is the worst in terms of explaining **medv**? [2 points] dis

2. For the simple linear regression using **dis** to predict **medv** (the top-right plot in sub-problem 1), let us look at the residual

$$resid_i = medv_i - fitted_i$$

against $fitted_i$ for each data point i.



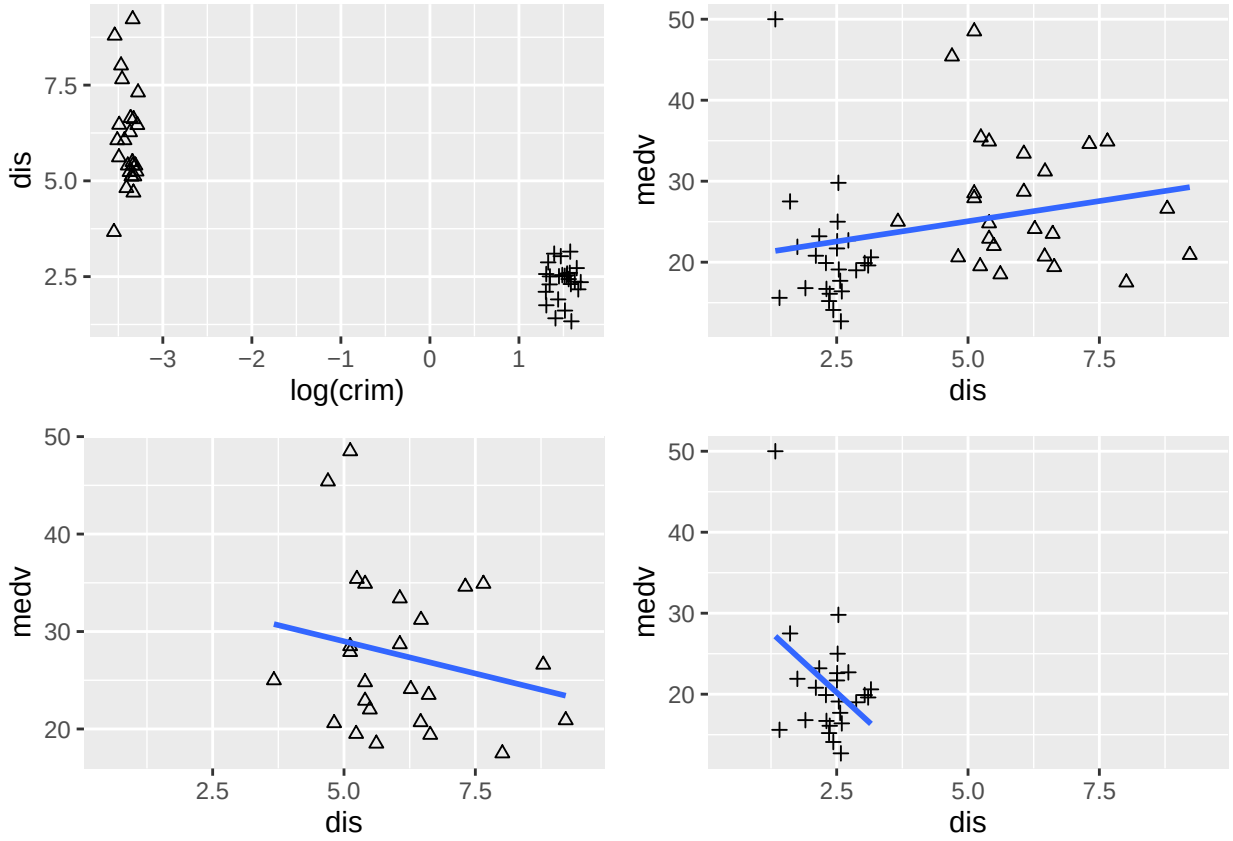
Suppose we run a new linear regression, based on the plot above, using $resid$ and $fitted$

$$resid = b_0 + b_1 \cdot fitted.$$

What is your best guess on b_0 and b_1? [3 points]

- (a) $b_0 < 0, b_1 < 0$
- (b) $b_0 > 0, b_1 = 0$
- (c) $b_0 = 0, b_1 = 0$ **Correct**
- (d) $b_0 = 0, b_1 < 0$
- (e) $b_0 = 0, b_1 > 0$

3. We want to understand the weak positive correlation between **dis** and **medv**, namely, why the further the distance is to the employment center, the more expensive the house is. To simplify the discussion, we look at the following subset of the data according to the variable **log(crim)**, which quantifies the per capita crime rate. In the following plots, '△' denotes data with a low crime rate, and '+' indicates data with a high crime rate.



Choose the options that make sense, combining all the information above. (Hint: you can choose multiple.) [4 points]

- (a) The above plot shows when holding the crime rate fixed, houses that are far away from the Boston employment centers tend to be less expensive than houses that are close, on average.
- (b) The above plot shows that for houses with a high crime rate, it is likely that they are closer to Boston employment centers.
- (c) The above color plot shows that for houses that are very far away from the Boston employment centers, it is most likely they have a low crime rate.
- (d) The weak positive coefficient of **dis** in explaining **medv** is because variables **dis** and **log(crim)** are negatively correlated. It means that a larger distance usually correlates with a low crime rate. The low crime rate is the reason behind a higher price on average.

ABCD

What is this phenomenon called in statistics? [1 point]

lurking variable (confounding)

Problem 8: Envelope game. [10 points]

At the end of BUS41000 class, Professor Liang decides to reward Tom for his hard work, and how much reward he can get depends on his probability skills. Professor Liang places two checks (one check is \$30, the other is \$70) into two envelopes. Note Tom has no idea about the value of the checks.

1. First, Tom decides to pick one envelope randomly. How much is his reward, in expectation? [2 points]

$$30 \times 0.5 + 70 \times 0.5 = 50$$

2. Suppose that the rule is changed slightly: Tom is allowed to choose one envelope, open it, and review the value of the check. Then he can decide whether to stick with the opened envelope or to swap to the other one. Tom recalls that one could use a randomized strategy to win more money. Here is Tom's new strategy: he draws a random number X using R/Excel from a normal distribution $X \sim N(50, 10^2)$, then compares this X with the value of the check he just opened. He will only keep the check if its value is larger than X. Otherwise, he will swap to the other envelope. Using this strategy, how much is Tom's reward, in expectation? How much more money he is going to get compared to the sub-problem 1? [8 points]

`Suppose Y is value of check you first pick`

Table	$X > Y$	$X < Y$
$Y = 30$	70, with probability 0.5×0.975	30, with probability 0.5×0.025
$Y = 70$	30, with probability 0.5×0.025	70, with probability 0.5×0.975

So the expectation is

$$70 \times 0.5 \times 0.975 + 30 \times 0.5 \times 0.025 \times 2 = 69$$

and $69 - 50 = 19$

Problem 9: Your intuition about correlation. [5 points]

1. Consider three stocks: A, B and C. Suppose $Corr(A, B) = 0$, $Corr(A, C) = 0$, what is the possible range of $Corr(B, C)$? [1 points]

- (a) [-1, 1] **Correct**
- (b) 0
- (c) None of the above

2. Suppose $Corr(A, B) = 1$, $Corr(A, C) = -1$, what is the possible range of $Corr(B, C)$? [1 points]

- (a) -1 **Correct**
- (b) 1
- (c) None of the above

3. Suppose $Corr(A, B) = 0.5$, $Corr(A, C) = 0.5$, what is the possible range of $Corr(B, C)$? [3 points]

- (a) [0.5, 1]
- (b) [-0.5, 1] **Correct**
- (c) [-1, 0.5]
- (d) None of the above

Extra Page for Calculations