Problem Set 3

Matt Krasnow

Due Saturday, September 28, 2024 at 11:59pm

Problem set policies. *Please provide concise, clear answers for each question while making sure to fully explain your reasoning. For problems asking for calculations, show your work in addition to clearly indicating the final answer. For problems involving `R`, be sure to include the code and output in your solution.*

Please submit the PDF of your knit solutions to Gradescope and be sure to assign which pages of your solution correspond to each problem. Make sure that the PDF is fully readable to the graders; e.g., make sure that lines don't run off the page margin.

We encourage you to discuss problems with other students (and, of course, with the teaching team), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution. Be aware that simply copying answers found online, whether human-generated or machine-generated, is a violation of the Honor Code.

Question 1: The Gauss Markov Theorem

In Lecture 5, we introduced the Gauss Markov Theorem, which states that, under the very general ELIH conditions, the OLS estimates are optimal among the class of linear unbiased estimators. By optimal, we mean they have minimum variance in the class (we call them **BLUE**: **b**est **l**inear **u**nbiased **e**stimators). In this problem, you will prove the Gauss Markov Theorem in the simple linear regression setting. That is, we assume:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} f(0, \sigma^2) \qquad i = 1, \dots n$$

Note that linear estimators are of the form:

$$\hat{\beta} = \sum_{i=1}^{n} c_i Y_i \tag{1}$$

where $c_1, \dots c_n$ are constants.

(a) What must be the variance of any linear estimator of the form above?

$$\hat{\beta} = \sum_{i=1}^{n} C_i \left[\beta_0 + \beta_1 X_i + \varepsilon_i\right]$$

$$= \sum_{i=1}^{n} C_i \beta_0 + \sum_{i=1}^{n} C_i \beta_1 X_i + \sum_{i=1}^{n} C_i \varepsilon_i$$

$$Var\left[\hat{\beta}\right] = Var\left[\sum_{i=1}^{n} C_i \beta_0\right] + Var\left[\sum_{i=1}^{n} C_i \beta_1 X_i\right] + Var\left[\sum_{i=1}^{n} C_i \varepsilon_i\right]$$

$$0 + 0 + \sum_{i=1}^{n} C_i^2 Var[\varepsilon_i]$$

$$Since\ \varepsilon_i\ are\ iid, \forall i \in [1, n],$$

$$Var[\hat{\beta}] = \sum_{i=1}^{n} C_i^2 \sigma^2$$

$$Var[\hat{\beta}] = \sigma^2 \sum_{i=1}^{n} C_i^2$$

(b) Derive two conditions on the $c_1, \dots c_n$ that are required for the estimate of β_0 to be unbiased.

For $\hat{\beta}_0$ to be unbiased:

$$E\left[\hat{\beta}_0\right] - \beta_0 = 0$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} C_i Y_i$$

$$E\left[\hat{\beta}_0\right] = \sum_{i=1}^{n} E[C_i Y_i] \text{ (by linearity)}$$

$$E\left[\hat{\beta}_0\right] = \sum_{i=1}^{n} C_i E\left[\beta_0 + \beta_1 X_i + \varepsilon_i\right]$$

$$= \sum_{i=1}^{n} C_i \beta_0 + C_i \beta_1 X_i$$

$$= \beta_0 \sum_{i=1}^{n} C_i + \beta_1 \sum_{i=1}^{n} C_i X_i$$

For this to be unbiased:

$$\sum_{i=1}^{n} C_i = 1 \quad \text{and} \quad \sum_{i=1}^{n} C_i X_i = 0$$

(c) Derive two conditions on the $c_1, \dots c_n$ that are required for the estimate of β_1 to be unbiased.

Following a similar process:

$$E\left[\hat{\beta}_1\right] = \sum_{i=1}^{n} E\left[C_i Y_i\right]$$

$$= \sum_{i=1}^{n} C_i \left[E(\beta_0) + E(\beta_1 X_i) + E(\varepsilon_i)\right]$$

$$E\left[\hat{\beta}_1\right] = \beta_0 \sum_{i=1}^{n} C_i + \beta_1 \sum_{i=1}^{n} C_i X_i$$

$$E\left[\hat{\beta}_1\right] = \beta_1 \quad \text{under the following conditions:}$$

$$1) \quad \sum_{i=1}^{n} C_i = 0$$

$$2) \quad \sum_{i=1}^{n} C_i X_i = 1$$

(d) Show that the OLS estimators for β_0 and β_1 are in the class of linear unbiased estimators. That is, prove they are unbiased, and that they take the form in Equation 1.

We need to show that:

1) The estimators are linear combinations of Y_i

2) Show that the estimators are unbiased.

We know that:

$$\hat{\beta_0} = \bar{Y} - \hat{\beta_1}\bar{X}$$

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$$

$$= \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right) Y_i - \bar{Y}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$$

$$= \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right) Y_i}{S_{XX}}$$

$$\text{Let } C_i = \frac{X_i - \bar{X}}{S_{XX}}$$

This makes the statement true:

$$\hat{\beta_1} = \sum_{i=1}^{n} C_i Y_i \quad \text{for} \quad C_i = \frac{X_i - \bar{X}}{S_{XX}}$$

Now solving for $\hat{\beta_0}$:

$$\hat{\beta_0} = \bar{Y} - \hat{\beta_1}\bar{X}$$

$$= \frac{1}{n}\sum_{i=1}^{n} Y_i - \left(\sum_{i=1}^{n} C_{i1} Y_i\right)\bar{X}$$

$$= \sum_{i=1}^{n} \frac{1}{n} Y_i - \sum_{i=1}^{n} C_{i1} Y_i \bar{X}$$

$$= \sum_{i=1}^{n} Y_i \left[\frac{1}{n} - C_{i1} \bar{X} \right]$$

$$\text{Let } C_{i2} = \frac{1}{n} - C_{i1} \bar{X}$$

Thus:

$$\hat{\beta}_0 = \sum_{i=1}^{n} C_{i2} Y_i$$

(2) Showing that they are unbiased. Using the previous results, for $\hat{\beta}_1$:

$$1) \quad \sum_{i=1}^{n} C_i = 0$$

$$2) \quad \sum_{i=1}^{n} C_i x_i = 1$$

For $\hat{\beta}_0$:

$$1) \quad \sum_{i=1}^{n} C_i \text{ must equal } 1$$

$$2) \quad \sum_{i=1}^{n} C_i x_i \text{ must equal } 0$$

$$\text{Where } C_i = \frac{x_i - \bar{x}}{S_{xx}}$$

$$\text{Show: } \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 0$$

$$\sum_{i=1}^{n} \frac{x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} - \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 0$$

$$\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} - \frac{n\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 0$$

$$\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} - \frac{n\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = 0$$

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

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

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

$$= \frac{1}{S_{xx}} S_{xx} = 1$$

Thus, it is shown that:

$$\hat{\beta}_1 = \sum_{i=1}^{n} C_j Y_j \text{ is an unbiased estimator for } \beta_1, \text{ and is a linear unbiased estimator.}$$

Now for $\hat{\beta}_0$:

Show:

$$\sum_{i=1}^{n} C_i = 1$$

$$\sum_{i=1}^{n} x_i C_i = 0$$

$$C_i = \frac{1}{n} - \frac{x_i - \bar{x}}{S_{xx}} \cdot \bar{x}$$

$$\sum_{i=1}^{n}\left(\frac{1}{n} - \frac{x_i-\bar{x}}{S_{xx}} \cdot \bar{x}\right) = 1 - \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{S_{xx}} \cdot \bar{x}\right)$$

$$1 - \left(\frac{\bar{x}}{S_{xx}}\sum_{i=1}^{n} x_i - \frac{n\bar{x}^2}{S_{xx}}\right) = 1$$

$$\sum_{i=1}^{n} x_i = n\bar{x}, \text{ thus:}$$

$$1 - \left(\frac{n\bar{x}^2}{S_{xx}} - \frac{n\bar{x}^2}{S_{xx}} \right) = 1 - 0 = 1$$

$$\text{Thus,} \sum_{i=1}^{n} C_{i2} = 1$$

Now showing:

$$\sum_{i=1}^{n} C_{i2} x_i = 0$$

$$\sum_{i=1}^{n} \left(\frac{1}{n} - \frac{x_i - \bar{x}}{S_{xx}} \right) x_i = 1 - \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}} \right) x_i$$

As previously mentioned:

$$\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}} \right) x_i = 1$$

Thus:

$$1 - 1 = 0$$

$$\text{Thus,} \sum_{i=1}^{n} C_{i2} x_i = 0$$

Since both conditions on the C_i are satisfied, they must be unbiased, as shown by the answers to parts b and c.

(e) Derive the variances of the OLS estimators of β_0 and β_1 and show they are of the form derived in part (a).

$$e) \text{ Derive variances of } \hat{\beta}_0 \text{ and } \hat{\beta}_1 \text{ for OLS.}$$

$$\text{For } \hat{\beta}_1 :$$

$$\hat{\beta}_1 = \sum_{i=1}^{n} C_{i1} y_i$$

$$\hat{\beta_1} = \sum_{i=1}^{n} C_{i1}(\beta_0 + \beta_1 X_i + \epsilon_i)$$

$$\hat{\beta_1} = \beta_0 \sum_{i=1}^{n} C_{i1} + \beta_1 \sum_{i=1}^{n} C_{i1} X_i + \sum_{i=1}^{n} C_{i1}\epsilon_i$$

Since we know $\sum_{i=1}^{n} C_{i1} = 0$ and $\sum_{i=1}^{n} C_{i1} X_i = 1$,

$$\hat{\beta_1} = \beta_1 + \sum_{i=1}^{n} C_{i1}\epsilon_i$$

Thus, $\text{Var}[\hat{\beta_1}] = 0 + \sum_{i=1}^{n} \text{Var}[C_{i1}\epsilon_i]$

$$= \sum_{i=1}^{n} C_{i1}^2 \sigma^2$$

$$= \sigma^2 \sum_{i=1}^{n} C_{i1}^2$$

$$= \sigma^2 \frac{1}{S_{xx}} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$= \sigma^2 \frac{1}{S_{xx}}$$

$$\boxed{\text{Var}[\hat{\beta_1}] = \frac{\sigma^2}{S_{xx}}}$$

Now for $\hat{\beta_0}$:

$$\hat{\beta_0} = \sum_{i=1}^{n} C_{i2} y_i$$

$$\hat{\beta_0} = \sum_{i=1}^{n} C_{i2}(\beta_0 + \beta_1 X_i + \epsilon_i)$$

$$\hat{\beta_0} = \beta_0 \sum_{i=1}^{n} C_{i2} + \beta_1 \sum_{i=1}^{n} C_{i2} X_i + \sum_{i=1}^{n} C_{i2}\epsilon_i$$

Since $\sum_{i=1}^{n} C_{i2} = 1$ and $\sum_{i=1}^{n} C_{i2} X_i = 0$,

$$\hat{\beta}_0 = \sum_{i=1}^{n} C_{i2} \epsilon_i$$

$$\text{Var}[\hat{\beta}_0] = \sum_{i=1}^{n} C_{i2}^2 \text{Var}[\epsilon_i]$$

$$= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n} - C_{i2} \bar{X} \right)^2$$

$$= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} - \frac{2C_{i2}\bar{X}}{n} + C_{i2}^2 \bar{X}^2 \right)$$

$$= \sigma^2 \left(\frac{1}{n} - 0 + \frac{\bar{X}^2}{S_{xx}} \right)$$

$$\boxed{\text{Var}[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} - \frac{\bar{X}^2}{S_{xx}} \right)}$$

Because $\sum_{i=1}^{n} C_{i2} = 0$ and $\sum_{i=1}^{n} C_{i2}^2 = \frac{1}{S_{xx}}$

Thus, $\text{Var}[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} - \frac{\bar{X}^2}{S_{xx}} \right)$

Final conclusion:

The variance of $\hat{\beta}_1$ is of the form $\sigma^2 \sum_{i=1}^{n} C_{i1}^2$ as found in part A.

The variance of $\hat{\beta}_0$ is of the form $\sigma^2 \sum_{i=1}^{n} C_{i2}^2$ as found in part A.

It is proven!

Question 2: Student height revisited

Question 1 on HW2 asked that you respond to a survey so that we could build our own dataset to explore. A subset of you responded by September 16, when I released the version of the dataset we used on that assignment. By the time the homework was due on September 20, more of you responded, but still not everyone.[1] I used the final (Sept. 20) cut of data to refit the simple linear model we introduced in class, where I regressed `studentheight` on `midparentheight`. However, this time `midparentheight` was defined as in HW2:

`midparentheight` = `paternalheight` + 1.08×`maternalheight`

The output from my linear regression is shown below:

```
Coefficients:
                Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)     10.42160    5.23422      1.991      0.0495 *
midparentheight 0.17181     0.03805      4.516      1.91e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.811 on 90 degrees of freedom
  (16 observations deleted due to missingness)
Multiple R-squared:  0.1847,    Adjusted R-squared:  0.1757
F-statistic: 20.39 on 1 and 90 DF,  p-value: 1.908e-05
```

(a) How many students responded to the survey?

$$DF = n - p$$

$$90 = n - 2$$

92 students, however 16 had missing data, so 108 students total responded.

(b) Interpret the R^2 from this model.

R^2 value of 0.1847 indicates that the model explains 18.47% of the variation in the data.

In other words, the model is 18% better than simply using the mean to predict.

In context of the problem, 18.47% of the variation in student height is predicted by the linear model with midpar

(c) Suppose that I had instead used the correct version of the independent variable:

Intuitively, this will NOT change the validity of the model—however, certain values would change:

Values that will change:

−Slope estimate and SE (both would double)

[1] Unfortunately, I didn't ask for your names, so I can't deduct points for not responding. . . I would have been ruthless.

Unchanged:

$-$Intercept & SE, scaling does not affect when $x = 0$

$-$T values, because all components are multiplied by the same factor.

$-$P values, no change in t

$-DF$

$-R^2$ (the fit of the model has not changed.)

Mathematical justification:

Slope doubles:

$$\beta_1^* = \frac{\beta_1}{0.5} = 2\beta_1$$

$$t = \frac{\beta_1^*}{SE[\beta_1^*]} = \frac{2\beta_1}{2SE[\beta_1]} = \frac{\beta_1}{SE[\beta_1]}$$

`midparentheight`$^* = \frac{1}{2}\times$ (`paternalheight` + 1.08×`maternalheight`)

and refit the model. What would change in the table above? Justify your answer mathematically.

(d) Now suppose instead that we had scaled our **dependent variable** by dividing it by 2. That is, we regressed:

`studentheight`$^* = \frac{1}{2} \times$ `studentheight`

on

`midparentheight` = `paternalheight` + 1.08×`maternalheight`

What would change in the table above? Justify your answer mathematically.

Values changed:

- Slope and SE

- Intercept and SE

Mathematical justification:

$$\beta_0^* = \frac{1}{2}\beta_0 = \frac{1}{2}\beta_0$$

$$SE[\beta_0^*] = SE[\frac{1}{2}\beta_0] = \frac{1}{2}SE[\beta_0]$$

$$\beta_1^* = \frac{1}{2}\beta_1 = \frac{1}{2}\beta_1$$

$$SE[\beta_1^*] = SE[\frac{1}{2}\beta_1] = \frac{1}{2}SE[\beta_1]$$

$$t = \frac{\beta_1^*}{SE[\beta_1^*]} = \frac{\frac{1}{2}\beta_1}{\frac{1}{2}SE[\beta_1]} = \frac{\beta_1}{SE[\beta_1]}$$

No change in t and DF implies p does not change.

R^2 does not change because the fit of the model has not changed.

Question 3: Confidence interval intuition

In Lecture 5, we developed the following confidence interval for $\mu_{Y|X=x_0}$ in simple regression:

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t^*_{n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

where t^*_{n-2} represents an appropriate quantile from a t-distribution with $n-2$ degrees of freedom.

(a) It should be clear mathematically that this confidence is narrowest at $\bar{x}$. Provide some intuition about why this is the case. Do not ask the TF's for intuition.

It makes sense mathematically
because the standard error gets smaller because

$$\lim_{x_0 \to \bar{x}} (x_0 - \bar{x})^2 = 0$$

The intuition for this can be expressed as follows: Standard error is a function of squared distance from the mean. So it
follows intuitively that when you get farther from the mean, the error will likely go up.

Also, we are estimating the mean parameter – so, it makes sense that our estimates that are close to the thing we are trying to estimate will
be better.

(b) An approximate confidence interval that holds for x_0 near $\bar{x}$ is:

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t^* \times \frac{\hat{\sigma}}{\sqrt{n}}$$

Provide intuition for why this approximation holds.

$$\textit{First I will demonstrate why this is the case mathematically},\ \textit{Since}\ \lim_{x_0 \to \bar{x}} \frac{(x_0 - \bar{x})^2}{S_{xx}} = 0,$$

$$\lim_{x_0 \to \bar{x}} \left(\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t^*_{n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right) =$$

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t^*_{n-2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$

Now, providing intuition for this:

If we assume that x_0 is close to bar{x}, this means that the distribution of x_0 is very closely centered around bar{x}. This means that it varies less, and since the variance tends to be less, we can approximate the regression line to have less variation, as demonstrated by the smaller variability term.

Question 4: Simulation to explore violations of assumptions

In this problem we will use simulation to explore the *robustness* of inference from linear regression. *Robustness* refers to the quality of the performance in the presence of assumption violations. Specifically, we will explore how the Type I error rate for our test of the hypothesis: $H_0 : \beta_1 = 0$ is affected when the normality and constant variance assumptions are violated.

To simulate data under the null (i.e., with $\beta_1 = 0$), we will use the following model:

$$Y_i = 10 + 0 \cdot x_i + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} f(0, \sigma^2) \qquad i = 1, \dots n$$

and we will use three different distributions for $f()$:

(i) $\varepsilon_i \sim N(\mu = 0, \sigma^2 = 4)$
(ii) $\varepsilon_i \sim a + b \cdot \text{Exp}(\alpha = 1)$
(iii) $\varepsilon_i | X_i = x_i \sim w \cdot N(0, x_i^2)$

and then we will fit a simple linear regression (with both an intercept and a slope) with `lm()`. We will also vary $n \in \{10, 30\}$.

Throughout, we will simulate the independent variable from a Uniform(0,1) distribution. That is:

$$X_i \sim \text{Unif}(\min = 0, \max = 1) \qquad i = 1, \dots, n$$

(a) Determine the values of a, b, and w above so that the residuals have a mean of zero and variance of 4 marginally for conditions (ii) and (iii). Be sure to show your work.

$$e_i \sim a + b\exp(\alpha = 1)$$

$$\mathbb{E}[e_i] = 1$$

$$\text{Var}[e_i] = 1$$

$$\mathbb{E}[G_i] = a + b$$

$$\text{Thus, } a = -b$$

$$\text{Var}[G_i] = y = 0.4b^2$$

$$b = \pm 2$$

$$a = \mp 2$$

$$\text{If } b = 2, a = -2$$

$$\text{If } b = -2, a = 2$$

$$\text{For convention, let's use } a = -2 \text{ and } b = 2$$

$$X_i \sim \text{unif}(0, 1)$$

$$\mathbb{E}[G_i|X_i] = 0$$

$$X_i \text{ is i.i.d.}$$

$$\mathbb{E}[G_i] = \mathbb{E}_{X_i}\left[\mathbb{E}\left[G_i|X_i\right]\right] = 0$$

$$\text{Var}[G_i] = \mathbb{E}_{X_i}\left[\text{Var}[G_i|X_i]\right]$$

$$= \mathbb{E}_{X_i}\left[w^2 X_i^2\right] = w^2\mathbb{E}[X_i^2]$$

$$\text{Since } X_i \sim \text{unif}(0, 1)$$

$$\mathbb{E}[X_i^2] = \int_0^1 x^2 dx = \frac{1}{3}$$

$$\text{For Var}[G_i] = y \Rightarrow w^2 \cdot \frac{1}{3} = 4$$

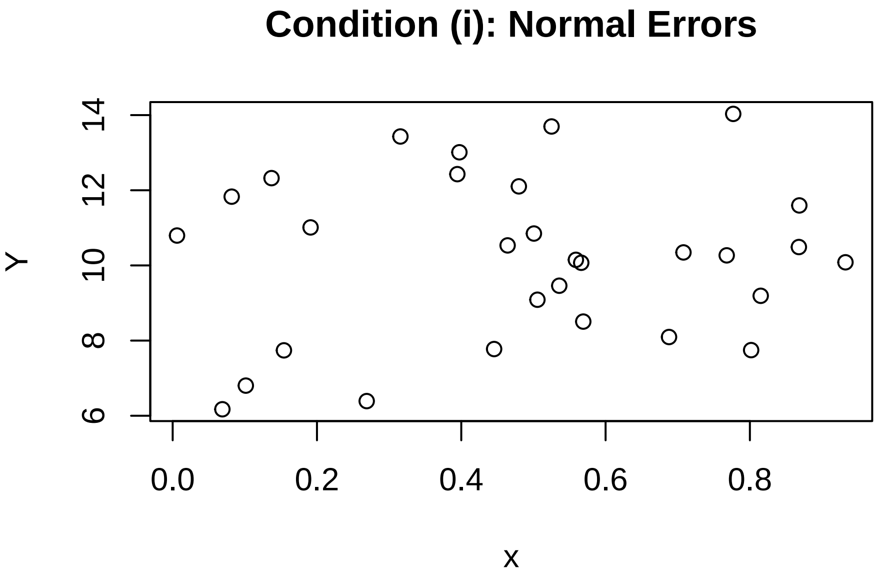
$$w = \pm 2\sqrt{3}$$

$$w = 2\sqrt{3}$$

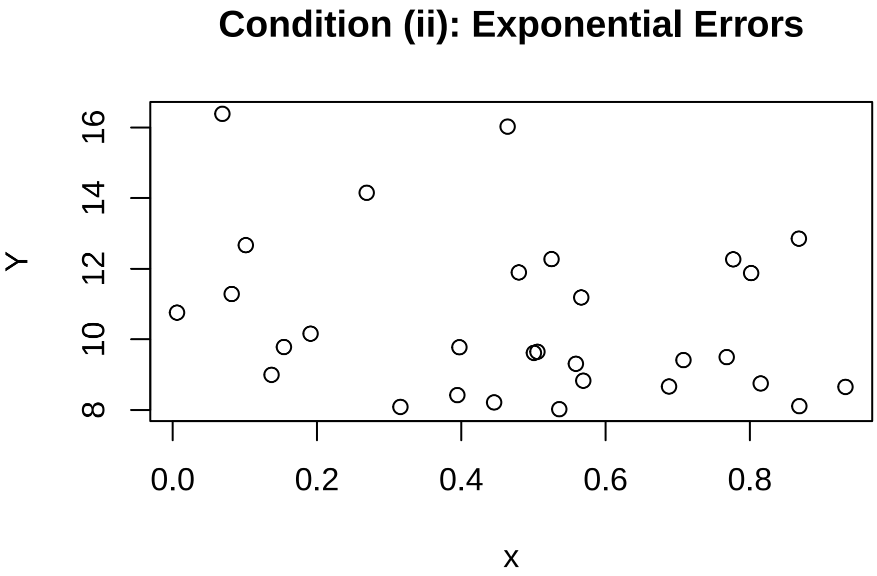
$$\text{Thus, } a = -2, b = 2, w = 2\sqrt{3}$$

(b) Create three samples of $n = 30$ observations: one for each of the three residual conditions above. The partial code given below in part (c) below may help you organize your thoughts. Create a scatterplot of $\mathbf{Y}$ versus $\mathbf{x}$ for each of the three conditions. Comment on what you see in relationship to the regression assumptions.

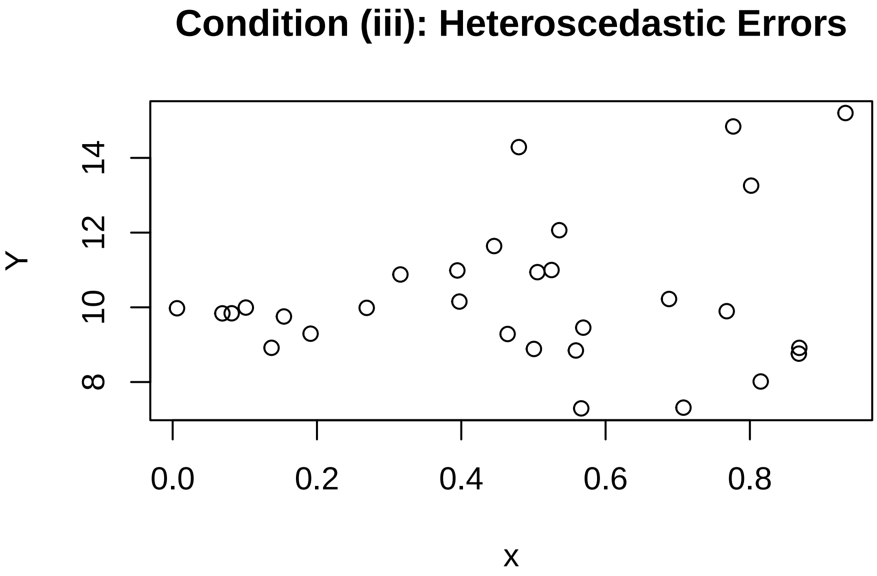
```r
set.seed(139)

n <- 30

# Generate X ~ Uniform(0,1)
x <- runif(n, min=0, max=1)

# Condition (i): epsilon ~ N(0,4)
epsilon_i <- rnorm(n, mean=0, sd=2)
Y_i <- 10 + 0 * x + epsilon_i
# Plot Y vs x
plot(x, Y_i, main="Condition (i): Normal Errors", xlab="x", ylab="Y")
```



```r
# Condition (ii): epsilon ~ a + b * Exp(1), with a = -2, b = 2
a <- -2
b <- 2
epsilon_ii <- a + b * rexp(n, rate=1)
Y_ii <- 10 + 0 * x + epsilon_ii
# Plot Y vs x
plot(x, Y_ii, main="Condition (ii): Exponential Errors", xlab="x", ylab="Y")
```



```r
# Condition (iii): epsilon | x ~ w * N(0, x^2), with w = 2 * sqrt(3)
w <- 2 * sqrt(3)
epsilon_iii <- w * rnorm(n, mean=0, sd=x)
Y_iii <- 10 + 0 * x + epsilon_iii
# Plot Y vs x
plot(x, Y_iii, main="Condition (iii): Heteroscedastic Errors", xlab="x", ylab="Y")
```



Condition (i): The scatterplot shows no apparent relationship between Y and x, with data points evenly spread, indicating homoscedasticity and normality of residuals. This meets the regression assumptions.

Condition (ii): The scatterplot reveals a skewed distribution of Y values due to the exponential errors, violating the normality assumption. However, the variance appears constant, so the homoscedasticity assumption holds.

Condition (iii): The scatterplot exhibits a fan-shaped pattern where the spread of Y increases with x, indicating heteroscedasticity. This violates the constant variance assumption, though the residuals are normally distributed conditional on x.

(c) Define a function called `reg.sim` that could be used to perform the simulation for the specific scenarios above. It should have at least the following arguments:

- `n`: the sample size, with a default of 10.
- `e.dist`: a string/character variable with 3 options: 'norm', 'expo', or 'weighted' corresponding to the three conditions above, with a default of 'norm'.
- `nsims`: the number of simulation repetitions, with a default of 2000.
- `seed`: the value for `set.seed`, with a default of 'NA'.

and return a list that includes:

- `beta`: the OLS estimate for $\boldsymbol{\beta}_1$
- `t.pvalue`: the p-value associated with the OLS estimate for $\boldsymbol{\beta}_1$ estimate

The code below frames the function for you. Your job is to fill out the meat of the function, and then use it in the next part.

```r
reg.sim = function(n=10,e.dist="norm",nsims=2000,seed=NA){
  if(!is.na(seed)){set.seed(seed)}
  # your code here
  # 0. before the for loop define storage vectors for saving betas and t.pvalues
  for(i in 1:nsims){
    # Create 'nsims' sets of data:
    # 1. generate the data, first x~Unif(0,1) then y based on the scenario
    if(e.dist=="norm"){
      # do something using rnorm()
    }
    if(e.dist=="expo"){
      # do something using rexp()
    }
    if(e.dist=="weighted"){
      # do something using rnorm()
    }
    # 2. calculate OLS estimates and store 2 things (fine to use summary from lm):
    #     a. the estimate for the slope
    #     b. the standard p-value from OLS
    if(i%%100==0){print(paste("done with iteration, i =",i))}
  }
  # 3. return the two variables as a list: beta and p-value
  return(list(beta=, pvalue=))
}
```

```r
reg.sim = function(n=10, e.dist="norm", nsims=2000, seed=NA){
  if(!is.na(seed)){set.seed(seed)}

  # 0. Define storage vectors for beta estimates and p-values
  beta_estimates <- numeric(nsims)
  p_values <- numeric(nsims)
  beta0 <- 10

  for(i in 1:nsims){
    # 1. Generate x ~ Uniform(0,1)
    x_i <- runif(n, min=0, max=1)

    # Generate epsilon based on the scenario
    if(e.dist=="norm"){
      epsilon_i <- rnorm(n, mean=0, sd=2)
    }
    if(e.dist=="expo"){
      a <- -2
      b <- 2
      epsilon_i <- a + b * rexp(n, rate=1)
    }
    if(e.dist=="weighted"){
      w <- 2 * sqrt(3)
      epsilon_i <- w * rnorm(n, mean=0, sd=x_i)
    }

    # Generate Y
    Y_i <- beta0 + 0 * x_i + epsilon_i

    # 2. Calculate OLS estimates and store beta and p-value
    model <- lm(Y_i ~ x_i)
    summary_model <- summary(model)
    beta_estimates[i] <- summary_model$coefficients["x_i","Estimate"]
    p_values[i] <- summary_model$coefficients["x_i","Pr(>|t|)"]

    if(i%%100==0){print(paste("done with iteration, i =",i))}
  }

  # 3. Return the beta estimates and p-values
  return(list(beta=beta_estimates, pvalue=p_values))
}
```

(d) Use your function above to perform the simulation with 2000 repetitions under the 6 scenarios - i.e., the combinations of the three residual conditions and the two sample sizes. Determine the rejection rate for each of the 6 scenarios, and present these results in an organized tabular form.

```r
# Parameters
nsims <- 2000
alpha_level <- 0.05
seed_value <- 123

# Initialize results data frame
results <- data.frame(
  n = integer(),
  e.dist = character(),
  rejection_rate = numeric(),
  stringsAsFactors=FALSE
)

# Sample sizes and error distributions
n_values <- c(10, 30)
e_dists <- c("norm", "expo", "weighted")

for (n in n_values){
  for (e.dist in e_dists){
    sim_result <- reg.sim(n=n, e.dist=e.dist, nsims=nsims, seed=seed_value)
    p_values <- sim_result$pvalue
    rejection_rate <- mean(p_values < alpha_level)
    # Store the results
    results <- rbind(results, data.frame(n=n, e.dist=e.dist, rejection_rate=rejection_rate))
  }
}
```

```
## [1] "done with iteration, i = 100"
## [1] "done with iteration, i = 200"
## [1] "done with iteration, i = 300"
## [1] "done with iteration, i = 400"
## [1] "done with iteration, i = 500"
## [1] "done with iteration, i = 600"
## [1] "done with iteration, i = 700"
## [1] "done with iteration, i = 800"
## [1] "done with iteration, i = 900"
## [1] "done with iteration, i = 1000"
## [1] "done with iteration, i = 1100"
## [1] "done with iteration, i = 1200"
## [1] "done with iteration, i = 1300"
## [1] "done with iteration, i = 1400"
## [1] "done with iteration, i = 1500"
## [1] "done with iteration, i = 1600"
## [1] "done with iteration, i = 1700"
## [1] "done with iteration, i = 1800"
## [1] "done with iteration, i = 1900"
## [1] "done with iteration, i = 2000"
## [1] "done with iteration, i = 100"
## [1] "done with iteration, i = 200"
## [1] "done with iteration, i = 300"
## [1] "done with iteration, i = 400"
## [1] "done with iteration, i = 500"
## [1] "done with iteration, i = 600"
## [1] "done with iteration, i = 700"
## [1] "done with iteration, i = 800"
## [1] "done with iteration, i = 900"
## [1] "done with iteration, i = 1000"
## [1] "done with iteration, i = 1100"
## [1] "done with iteration, i = 1200"
## [1] "done with iteration, i = 1300"
## [1] "done with iteration, i = 1400"
## [1] "done with iteration, i = 1500"
## [1] "done with iteration, i = 1600"
## [1] "done with iteration, i = 1700"
## [1] "done with iteration, i = 1800"
## [1] "done with iteration, i = 1900"
## [1] "done with iteration, i = 2000"
## [1] "done with iteration, i = 100"
## [1] "done with iteration, i = 200"
## [1] "done with iteration, i = 300"
## [1] "done with iteration, i = 400"
## [1] "done with iteration, i = 500"
## [1] "done with iteration, i = 600"
## [1] "done with iteration, i = 700"
## [1] "done with iteration, i = 800"
## [1] "done with iteration, i = 900"
## [1] "done with iteration, i = 1000"
## [1] "done with iteration, i = 1100"
## [1] "done with iteration, i = 1200"
## [1] "done with iteration, i = 1300"
## [1] "done with iteration, i = 1400"
## [1] "done with iteration, i = 1500"
## [1] "done with iteration, i = 1600"
## [1] "done with iteration, i = 1700"
## [1] "done with iteration, i = 1800"
## [1] "done with iteration, i = 1900"
## [1] "done with iteration, i = 2000"
## [1] "done with iteration, i = 100"
## [1] "done with iteration, i = 200"
## [1] "done with iteration, i = 300"
## [1] "done with iteration, i = 400"
## [1] "done with iteration, i = 500"
## [1] "done with iteration, i = 600"
## [1] "done with iteration, i = 700"
## [1] "done with iteration, i = 800"
## [1] "done with iteration, i = 900"
## [1] "done with iteration, i = 1000"
## [1] "done with iteration, i = 1100"
## [1] "done with iteration, i = 1200"
## [1] "done with iteration, i = 1300"
## [1] "done with iteration, i = 1400"
## [1] "done with iteration, i = 1500"
## [1] "done with iteration, i = 1600"
## [1] "done with iteration, i = 1700"
## [1] "done with iteration, i = 1800"
## [1] "done with iteration, i = 1900"
## [1] "done with iteration, i = 2000"
## [1] "done with iteration, i = 100"
## [1] "done with iteration, i = 200"
## [1] "done with iteration, i = 300"
## [1] "done with iteration, i = 400"
## [1] "done with iteration, i = 500"
## [1] "done with iteration, i = 600"
## [1] "done with iteration, i = 700"
## [1] "done with iteration, i = 800"
## [1] "done with iteration, i = 900"
## [1] "done with iteration, i = 1000"
## [1] "done with iteration, i = 1100"
## [1] "done with iteration, i = 1200"
## [1] "done with iteration, i = 1300"
## [1] "done with iteration, i = 1400"
## [1] "done with iteration, i = 1500"
## [1] "done with iteration, i = 1600"
## [1] "done with iteration, i = 1700"
## [1] "done with iteration, i = 1800"
## [1] "done with iteration, i = 1900"
## [1] "done with iteration, i = 2000"
## [1] "done with iteration, i = 100"
## [1] "done with iteration, i = 200"
## [1] "done with iteration, i = 300"
## [1] "done with iteration, i = 400"
## [1] "done with iteration, i = 500"
## [1] "done with iteration, i = 600"
## [1] "done with iteration, i = 700"
## [1] "done with iteration, i = 800"
## [1] "done with iteration, i = 900"
## [1] "done with iteration, i = 1000"
## [1] "done with iteration, i = 1100"
## [1] "done with iteration, i = 1200"
## [1] "done with iteration, i = 1300"
## [1] "done with iteration, i = 1400"
## [1] "done with iteration, i = 1500"
## [1] "done with iteration, i = 1600"
## [1] "done with iteration, i = 1700"
## [1] "done with iteration, i = 1800"
## [1] "done with iteration, i = 1900"
## [1] "done with iteration, i = 2000"
```

```r
# Present results in a table
print(results)
```

```
##    n   e.dist rejection_rate
## 1 10     norm         0.0480
## 2 10     expo         0.0390
## 3 10 weighted         0.0820
## 4 30     norm         0.0485
## 5 30     expo         0.0450
## 6 30 weighted         0.0790
```

(e) How does inference for the slope in simple regression behave in the presence of these assumption violations? To which assumption(s) is the inference more more robust? How does sample size play a role? Please limit your response to 5 or fewer sentences.

The normal errors scenario (condition i) serves as our baseline, with rejection rates close to the nominal 5% level, indicating proper Type I error control. For exponential errors (condition ii), the rejection rates remain relatively close to 5%, suggesting that inference is fairly robust to violations of the normality assumption. Heteroscedastic errors (condition iii) show the highest deviation from the 5% level, particularly for smaller sample sizes, indicating that inference is less robust to violations of the constant variance assumption. Increasing the sample size from 10 to 30 generally improves the performance across all conditions, bringing rejection rates closer to the nominal level, which demonstrates that larger sample sizes can help mitigate the impact of assumption violations. Overall, inference appears more robust to violations of normality than to violations of constant variance, with sample size playing a crucial role in improving the reliability of inference under various assumption violations.