POL SCI 231b (Spring 2017):

Problem Set 4 Solution Set

Prof. Thad Dunning/GSI Natalia Garbiras-Díaz

Dept. of Political Science

University of California, Berkeley

Suggested Solutions

Before completing this problem set, you should work Exercise Sets A and B in Freedman (2009, Chapter 4). You should also read and think about the discussion questions.

1. **Suppose that you have an $n \times 1$ column vector of observations on a variable Y and an $n \times 1$ column vector of observations on a variable X. You convert each variable to standard units. Then you run a regression of standardized Y on standardized X.**

   (a) **Show that the fitted slope coefficient on standardized X is r, the coefficient of correlation between Y and X.**

   Note that if we regress the (unstandardized) Y on an intercept and X, we can express the actual values in terms of the fitted values and the residuals: $Y_i = \hat{a} + \hat{b}X_i + e_i$, where $\hat{a}$ and $\hat{b}$ are the fitted intercept and slope coefficient, respectively. Thus, recalling that $\bar{e} = 0$, we have $\bar{Y} = \hat{a} + \hat{b}\bar{X}$, and so $Y_i - \bar{Y} = \hat{b}(X_i - \bar{X}) + e_i$. (Just subtract the expression for $\bar{Y}$ from the expression for Y_i). Then,

   $$\frac{Y_i - \bar{Y}}{S_Y} = \hat{b}\frac{S_X}{S_Y}\frac{X_i - \bar{X}}{S_X} + \frac{e_i}{S_Y}, \tag{1}$$

   where we have divided through by the standard deviation S_Y and then multiplied and divided $\hat{b}\frac{X_i - \bar{X}}{S_Y}$ by $\frac{S_X}{S_X}$. Now, we know that the fitted slope coefficient in the

unstandardized case is

$$\hat{b} = \frac{\text{cov}(X, Y)}{\text{var}(X)}. \tag{2}$$

So, the fitted coefficient on $\frac{X_i - \bar{X}}{S_X}$ in (1) is

$$\hat{b}\frac{S_X}{S_Y} = \frac{\text{cov}(X, Y)}{\text{var}(X)}\frac{S_X}{S_Y} = \frac{\text{cov}(X, Y)}{(S_X)(S_Y)} = r. \tag{3}$$

The key point: the slope coefficient in a standardized bivariate regression (that is, a regression of standardized Y on standardized X) is just r, the correlation coefficient between X and Y.

(b) **What assumptions of the linear regression model did you need for part (a)?** You don't need any assumptions of the linear regression model for part (a). This is a cold, mechanical fact about regression—it follows for any regression you fit of standardized Y on standardized X.

2. **Consider the linear regression model, $Y = X\beta + \epsilon$. The design matrix X is $n \times p$; the first column of X is all 1s. Make a list of the assumptions of the model.**

For the "usual" OLS regression model, the assumptions are:

(a) $Y = X\beta + \epsilon$.

(b) $\epsilon \perp\!\!\!\perp X$

(c) ϵ is i.i.d. with $\text{E}(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$.

An implicit assumption (readily verifiable from the data) is that X has full rank p. (Here, we may be considering fixed X or random X).

Some notes: (a) is the assumption that the model itself holds. Assumption (b) is independence of the error term and the design matrix. For (c), $\text{Var}(\epsilon_i) = \sigma^2$ is called "homoskedasticity" (or "homoscedasticity"). Together, $\epsilon \perp\!\!\!\perp X$ and $\text{E}(\epsilon_i) = 0$ imply $\text{E}(\epsilon_i|X) = 0$. The "classical" linear model adds $\epsilon_i \sim N(0, 1)$ to (c), i.e., the random variable ϵ_i is standard normal.

For other variants of the linear regression model such as GLS, the assumptions in (c) may vary. For example, we might have $\text{Var}(\epsilon_i) = \sigma_i$ (note the subscript on σ_i: each random variable ϵ_i has a possibly different variance idiosyncratic to i, which violates homoskedasticity). Or, we might have $\text{Cov}(\epsilon_i, \epsilon_j) \neq 0$. The usual OLS model rules out the latter case because each ϵ_i is independent (the first part of "i.i.d.")

3. **Now, suppose we fit the regression model in the previous question to data, giving $Y_i = X_i\hat{\beta} + e_i$. Here, $\hat{\beta} = (X'X)^{-1}X'Y$, and $e_i = Y_i - X_i\hat{\beta}$, where X_i is the ith row of X. Answer the following questions:**

(a) **True or false:** $\epsilon \perp X$ This is usually false: independence of the random variables ϵ_i and X_i does not translate to exact orthogonality of ϵ and X.

(b) **True or false:** $\epsilon \perp\!\!\!\perp X$

This is true if the model holds, by assmuption (b) in question 2.

(c) **True or false:** $e \perp\!\!\!\perp X$

Generally false. Since $e = Y - X\hat{\beta}$, e is a function of X, in particular, of $X\hat{\beta} = X(X'X)^{-1}X'Y$: indeed, $\hat{\beta}$ is constructed to minimize the sum of squared residuals, so e generally is not independent of X.

(d) **True or false:** $e \perp X$

True. This follows from the mechanics of regression and has nothing to do with the modeling assumptions.

(e) **Does $e \perp X$ help validate $\epsilon \perp\!\!\!\perp X$?**

No. As per the answer to (d), $e \perp X$ follows as a matter of algebra from any regression of Y on X; it does not validate the modeling assumption $\epsilon \perp\!\!\!\perp X$.

(f) **Does $\bar{e} = 0$ help validate $E(\epsilon_i) = 0$?**

No, $\bar{e} = 0$ follows mechanically in any regression as long as there is a constant column in X (i.e., there is an intercept). (In more detail, it follows from $e \perp X$ and thus $e'X = 0$, which implies $\sum_{i=1}^{n} e_i = 0$ and thus $\bar{e} = 0$ if the first column is constant). This does not validate the modeling assumption $E(\epsilon_i) = 0$.

(g) **Is $\sum_{i=1}^{n} \epsilon_i = 0$? Or is the sum typically around $\sigma\sqrt{n}$ in size?**

The sum is typically around $\sigma\sqrt{n}$. This follows from the central limit theorem (for the sum of independent random variables): the sampling distribution of ϵ is approximately normal (the approximation gets better as n increases).

4. **(Interaction terms). Scholars sometimes counsel the use of interaction models for testing conditional hypotheses. This exercise gives an example. Assume the following OLS model:**

$$Y_i = a + bX_i + cZ_i + dX_iZ_i + \epsilon_i \tag{4}$$

for all $i = 1, ..., n$. Here, X_iZ_i is the product of X_i and Z_i; this is usually called an "interaction term." The usual OLS assumptions apply, e.g., $\epsilon_i \perp\!\!\!\perp (X_i, Z_i, X_iZ_i)$ and the ϵ_i are i.i.d. with $E(\epsilon) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$ for all i.

(a) **What is the marginal effect of intervening to change X_i with Z_i held fixed? That is, if**

$$E(Y_i|X_i, Z_i) = a + bX_i + cZ_i + dX_iZ_i, \tag{5}$$

what is

$$\frac{\partial E(Y_i|X_i, Z_i)}{\partial X_i}? \tag{6}$$

And what is the marginal effect of intervening to change Z_i with X_i held fixed?

(Note: here, the "marginal effect" means, the change in the expected value of Y_i due to a one-unit change in X_i or Z_i).

The marginal effect of intervening to change X_i with Z_i held fixed is

$$\frac{\partial E(Y_i|X_i, Z_i)}{\partial X_i} = b + dZ_i, \tag{7}$$

while the marginal effect of intervening to change Z_i with X_i held fixed is

$$\frac{\partial E(Y_i|X_i, Z_i)}{\partial Z_i} = c + dX_i. \tag{8}$$

(b) **Now, suppose you run a regression of Y_i on a constant, X_i, Z_i, and X_iZ_i, obtaining**

$$\hat{Y}_i = \hat{a} + \hat{b}X_i + \hat{c}Z_i + \hat{d}X_iZ_i \tag{9}$$

for all i. Here, $\hat{Y}_i$ is the fitted ("predicted") value of Y_i, $\hat{a}$ is the fitted intercept, and $\hat{b}$, $\hat{c}$, and $\hat{d}$ are the fitted slope coefficients. Under the assumptions of the model in (4), $\hat{a}$ estimates a, $\hat{b}$ estimates b, and so on.

What is the estimated marginal effect of intervening to change X_i with Z_i held fixed? How about the estimated marginal effect of intervening to change Z_i with X_i held fixed?

Just replace the parameters in part (a) with the estimates. Then, the estimated marginal effect of intervening to change X_i with Z_i held fixed is

$$\frac{\partial \hat{Y}_i}{\partial X_i} = \hat{b} + \hat{d}Z_i, \tag{10}$$

while the estimated marginal effect of intervening to change Z_i with X_i held fixed is

$$\frac{\partial \hat{Y}_i}{\partial Z_i} = \hat{c} + \hat{d}X_i. \tag{11}$$

(c) **Let $\frac{\partial \hat{Y}_i}{\partial X_i}$ be the estimated marginal effect of intervening to change X_i with Z_i held fixed. Express the variance of $\frac{\partial \hat{Y}}{\partial X_i}$ in terms of variances and covariances of the variables and coefficient estimators in equation (9). (For convenience, treat X_i and Z_i as fixed, rather than as random variables).**

From (10),

$$\frac{\partial \hat{Y}_i}{\partial X_i} = \hat{b} + \hat{d}Z_i. \tag{12}$$

Thus,

$$\begin{aligned} \text{var}(\frac{\partial \hat{Y}_i}{\partial X_i}) &= \text{var}(\hat{b} + \hat{d}Z_i) \\ &= \text{var}(\hat{b}) + Z_i^2\text{var}(\hat{d}) + 2Z_i\text{cov}(\hat{b}, \hat{d}). \end{aligned} \tag{13}$$

where the second line of (13) follows from distributing the variance; here, Z_i factors out, because Z_i is fixed. If Z_i and X_i were random, we would condition on Z_i and X_i when taking the variance in (13).

(d) **Suppose that after fitting equation (9), you find that**

$$\hat{Y}_i = 1.2 + 2.3X_i + 0.5Z_i - 2.1X_iZ_i. \tag{14}$$

Moreover, the estimated variance-covariance matrix of $\hat{\beta} = (\hat{a}\,\hat{b}\,\hat{c}\,\hat{d})'$ is given by

$$\widehat{\text{cov}}(\hat{\beta}|X, Z, XZ) = \begin{pmatrix} 0.5 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0.9 & 0.5 & -0.3 \\ 0.4 & 0.5 & 0.4 & 0.3 \\ 0.7 & -0.3 & 0.3 & 0.7 \end{pmatrix} \tag{15}$$

Here, n is large, so the sampling distribution of $\hat{\beta}$ is approximately normal.

i. **Conduct a test of the null hypothesis that $b = 0$.**
The estimated variance of $\hat{b}$ is the (2,2) element of the covariance matrix $\text{cov}(\hat{\beta}|X, Z, XZ)$. This is 0.9, so the estimated standard error is the square root: $\sqrt{0.9} \doteq 0.95$. The t-statistic is the absolute value of the coefficient estimate $\hat{b}$ divided by the estimated standard error:

$$t = |\frac{2.3}{0.95}| = 2.42 \tag{16}$$

So the t-test suggests that $\hat{b}$ is highly significant: we can reject the null hypothesis that $b = 0$.

ii. **Conduct a test of the null hypothesis that $d = 0$.**
The estimated variance of $\hat{d}$ is the (4,4) element of $\widehat{\text{cov}}(\hat{\beta}|X, Z, XZ)$. This is 0.7, so the estimated standard error is the square root: $\sqrt{0.7} \doteq 0.84$. The t-statistic is the absolute value of $\hat{d}$ divided by the estimated standard error:

$$t = |\frac{-2.1}{0.84}| = 2.5. \tag{17}$$

So the t-test suggests that $\hat{d}$ is highly significant: we can reject the null hypothesis that $d = 0$.

iii. **Conduct a test of the null hypothesis that the marginal effect of intervening to change X_i, with Z_i held fixed at $Z_i = 1$, is zero.**
The estimated marginal effect of intervening to change X_i is $\hat{b} + \hat{d}Z_i$ (see b above). So when $Z_i = 1$, the estimated marginal effect is $\hat{b}+\hat{d} = 2.3-2.1 = 0.2$. From (c), the variance of the estimated marginal effect is $\text{var}(\hat{b}) + Z_i^2\text{var}(\hat{d}) + 2Z_i\text{cov}(\hat{b}, \hat{d})$. Using the (2,2), (4,4), and (4,2)—or equivalently, the (2,4)—

elements of $\widehat{\text{cov}}(\hat{\beta}|X, Z, XZ)$, we have

$$\begin{aligned} \text{var}(\hat{b}) + Z_i^2\text{var}(\hat{d}) + 2Z_i\text{cov}(\hat{b}, \hat{d}) &= 0.9 + Z_i^2 0.7 - 2Z_i 0.3 \\ &= 0.9 + 0.7 - 2(0.3) \\ &= 1.0 \end{aligned} \tag{18}$$

where the second line follows from $Z_i = 1$. Thus, the estimated standard error is $\sqrt{1.0} = 1.0$. Finally, the t-statistic is the absolute value of the estimated marginal effect over the estimated standard error:

$$t = |\frac{0.2}{1.0}| = 0.2. \tag{19}$$

Thus, the estimated marginal effect is statistically insignificant: we cannot reject the null hypothesis that the marginal effect on Y_i of intervening to change X_i with Z_i held fixed at $Z_i = 1$ is zero.

iv. **Conduct a test of the null hypothesis that the marginal effect of intervening to change X_i, with Z_i held fixed at $Z_i = 10$, is zero.**
As before, the estimated marginal effect of intervening to change X_i is $\hat{b}+\hat{d}Z_i$, so when $Z_i = 10$, the estimated marginal effect is $\hat{b}+10\hat{d} = 2.3-10(2.1) = -18.7$. Again, the variance of the estimated marginal effect is $\text{var}(\hat{b}) + Z_i^2\text{var}(\hat{d}) + 2Z_i\text{cov}(\hat{b}, \hat{d})$. Here, with $Z_i = 10$, we have

$$\begin{aligned} \text{var}(\hat{b}) + Z_i^2\text{var}(\hat{d}) + 2Z_i\text{cov}(\hat{b}, \hat{d}) &= 0.9 + Z_i^2 0.7 - 2Z_i 0.3 \\ &= 0.9 + (10)^2 0.7 - 2(10)(0.3) \\ &= 0.9 + 70 - 6 \\ &= 64.9 \end{aligned} \tag{20}$$

Thus, the standard error is $\sqrt{64.9} \doteq 8.06$. Finally, the t-statistic is the absolute value of the estimated marginal effect over the estimated standard error:

$$t = |\frac{-18.7}{8.06}| = 2.32 \tag{21}$$

Thus, the estimated marginal effect is highly significant: we can reject the null hypothesis that the marginal effect on Y_i of intervening to change X_i, with Z_i held fixed at $Z_i = 10$, is zero.

v. **Comment on your results. Do they suggest that—given the model—the effect of X_i is conditional on Z_i?**
The estimated coefficients $\hat{b}$ and $\hat{d}$ are both highly significant. However, this does not imply that the marginal effect of intervening to change X_i is different from zero. Indeed, this depends on the level of Z_i. When Z_i is negative (or small and positive), the overall marginal effect may be positive (or zero). When Z_i is large and positive, however, the estimated marginal effect is significantly negative. Thus, if the model is correct, the results do suggest that the effect

of X_i is conditional on Z_i.

5. **(Function). Extend the OLS function you wrote for problem set 3 to report the estimated standard errors of $\widehat{\beta}$ and the corresponding t statistic and p-values from a two-sided test. Include a test comparing the results of your function to using the lm() command.**

```
ols <- function(y, X){

    X <- cbind(1, X)

    # getting the coefficients
    coefs <- solve(t(X)%*%X) %*% (t(X)%*%y)

    # and now for the SEs...
    # residuals
    e <- y - X %*% coefs
    # estimated sigma squared
    hat_sigma2 <- sum(t(e)%*%e) / (nrow(X)-length(coefs))
    # and we can find estimated standard errors from $\hat{\sigma^2}(X'X)^{-1}$.
    varcovmat <- hat_sigma2 * solve((t(X)%*%X))
    SEs <- sqrt(diag(varcovmat))

    # t-stats
    tstat <- coefs/SEs

    # p-values
    df <- nrow(X) - ncol(X)
    pvals <- pt(abs(tstat), df=df, ncp=0, lower.tail=FALSE) * 2


    # output
    out <- round(as.matrix(cbind(coefs, SEs, tstat, pvals)), 5)
    colnames(out) <- c("coefficient", "SEs", "t-statistic", "p-values")

    return(out)
}


# Testing the function
# Fake data
library(MASS)
sigma <- matrix(c(1, 0.5, 0, 0.5, 1, 0.5, 0, 0.5, 1), 3, 3)
X <- mvrnorm(n=100, mu=c(0, 0, 0), Sigma=sigma)
y <- apply(X, 1, sum) + rnorm(100, mean=0, sd=2)
```

```r
ols(y, X)
```

```
##      coefficient     SEs t-statistic p-values
## [1,]    -0.03277 0.21240    -0.15427  0.87772
## [2,]     0.60040 0.29566     2.03073  0.04505
## [3,]     1.04159 0.31266     3.33139  0.00123
## [4,]     1.01826 0.27693     3.67699  0.00039
```

```r
summary(lm(y ~ X))
```

```
##
## Call:
## lm(formula = y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7490 -1.5354 -0.1346  1.2199  5.1223
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03277    0.21240  -0.154 0.877720
## X1           0.60040    0.29566   2.031 0.045047 *
## X2           1.04159    0.31266   3.331 0.001228 **
## X3           1.01826    0.27693   3.677 0.000389 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.104 on 96 degrees of freedom
## Multiple R-squared:  0.4564,Adjusted R-squared:  0.4394
## F-statistic: 26.86 on 3 and 96 DF,  p-value: 1.062e-12
```

6. **(Simulation). Sample 100 units from a normal distribution with mean 2 and variance 4. Call this vector x1, which will be fixed (will not vary across simulations). You will conduct a simulation with 10,000 replicates. For each replicate:**

   - **Sample 100 realizations of i.i.d. ϵ from a normal distribution with mean 0 and** $\text{sd} = \sqrt{2}$ **and use that to construct Y as follows:** $Y_i = 1.5 + 3x_{1i} + \epsilon_i$**.**
   - **Fit a regression of Y on X_{100x2}, where each row of X is $[1 x_{1i}]$ for $i = 1, \ldots, 100$.**
   - **Save $\widehat{\beta}$, $\widehat{\text{SE}}(\widehat{\beta})$, and the t-statistic.**

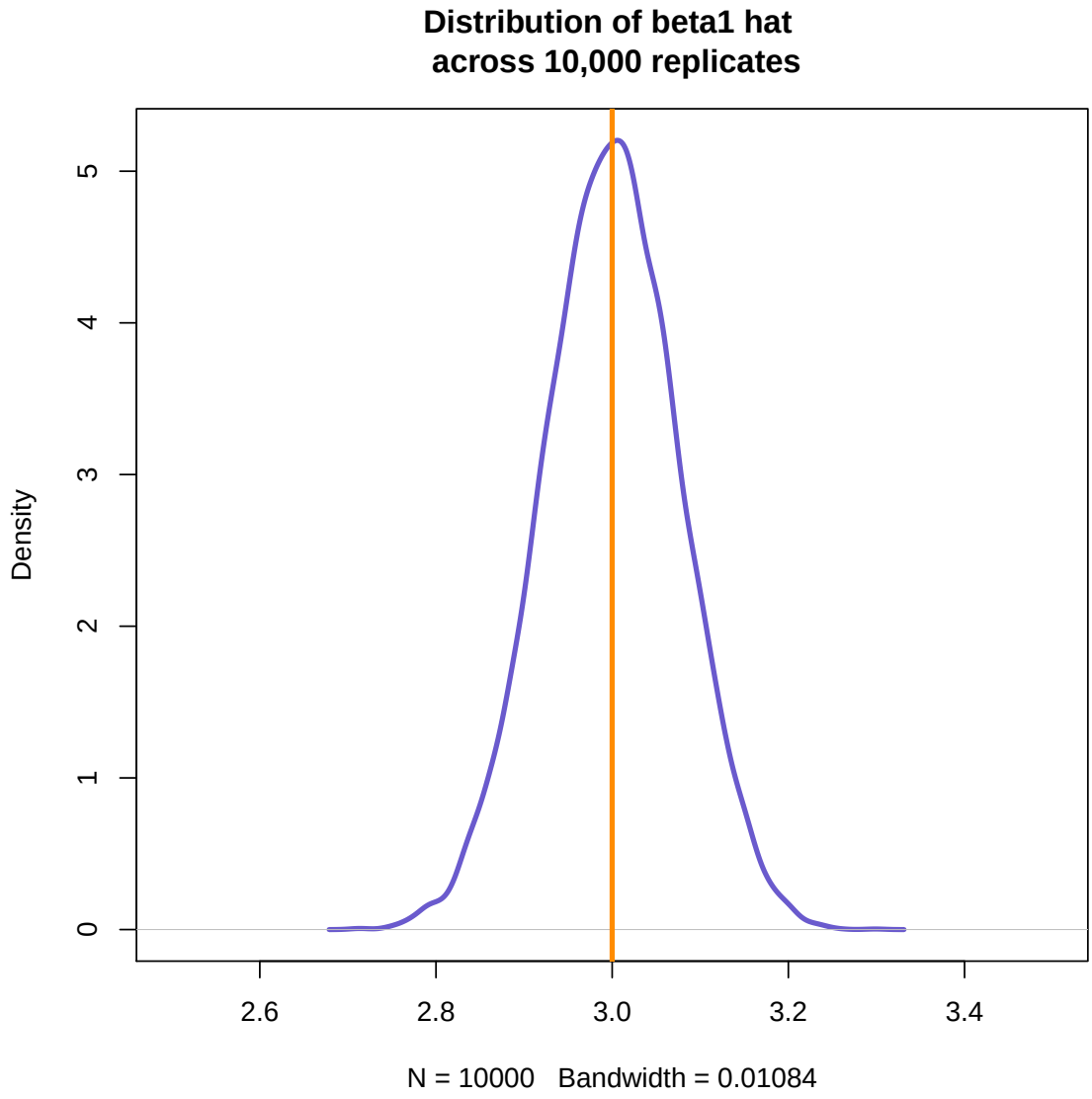
   Now,

(a) **Plot the distribution of $\widehat{\beta}$ across the 10,000 replicates. Is $\widehat{\beta}$ unbiased? Justify your answer.**

(b) **Calculate the standard deviation of the distribution in (a). How does this s.d. compare to the average of the $\widehat{\mathrm{SE}}(\widehat{\beta})$s across the 10,000 replicates? What does your answer indicate? And what is the theoretical standard error of $\widehat{\beta}$ in this simulation (i.e. based on $\sigma^2[X'X]^{-1}$)?**

(c) **Plot the distribution of the t-statistics. How close is the theoretical distribution of t to normal? Superimpose a normal curve on the plot (with same mean as t).**

```
x1 <- rnorm(100, mean=2, sd=2)
sigma2 <- 2

sim <- function(){

    epsilon <- rnorm(100, mean=0, sd=sqrt(sigma2))
    y <- 1.5 + 3*x1 + epsilon

    return(ols(y, x1)[2, 1:3])

}

results <- replicate(10000, sim())

par(mfrow=c(1,1))
plot(density(results[1,]), lwd=3, col="slateblue", xlim=c(2.5,3.5),
     main="Distribution of beta1 hat \n across 10,000 replicates")
abline(v=3, col="darkorange", lwd=3)
```



Beta hat is unbiased: the distribution of the estimator is centered around the true value of the parameter.

```
# sd of the distribution of beta1 hat
sd(results[1,])

## [1] 0.07600807

# average of SE(beta1 hat)
mean(results[2,])

## [1] 0.07555394
```

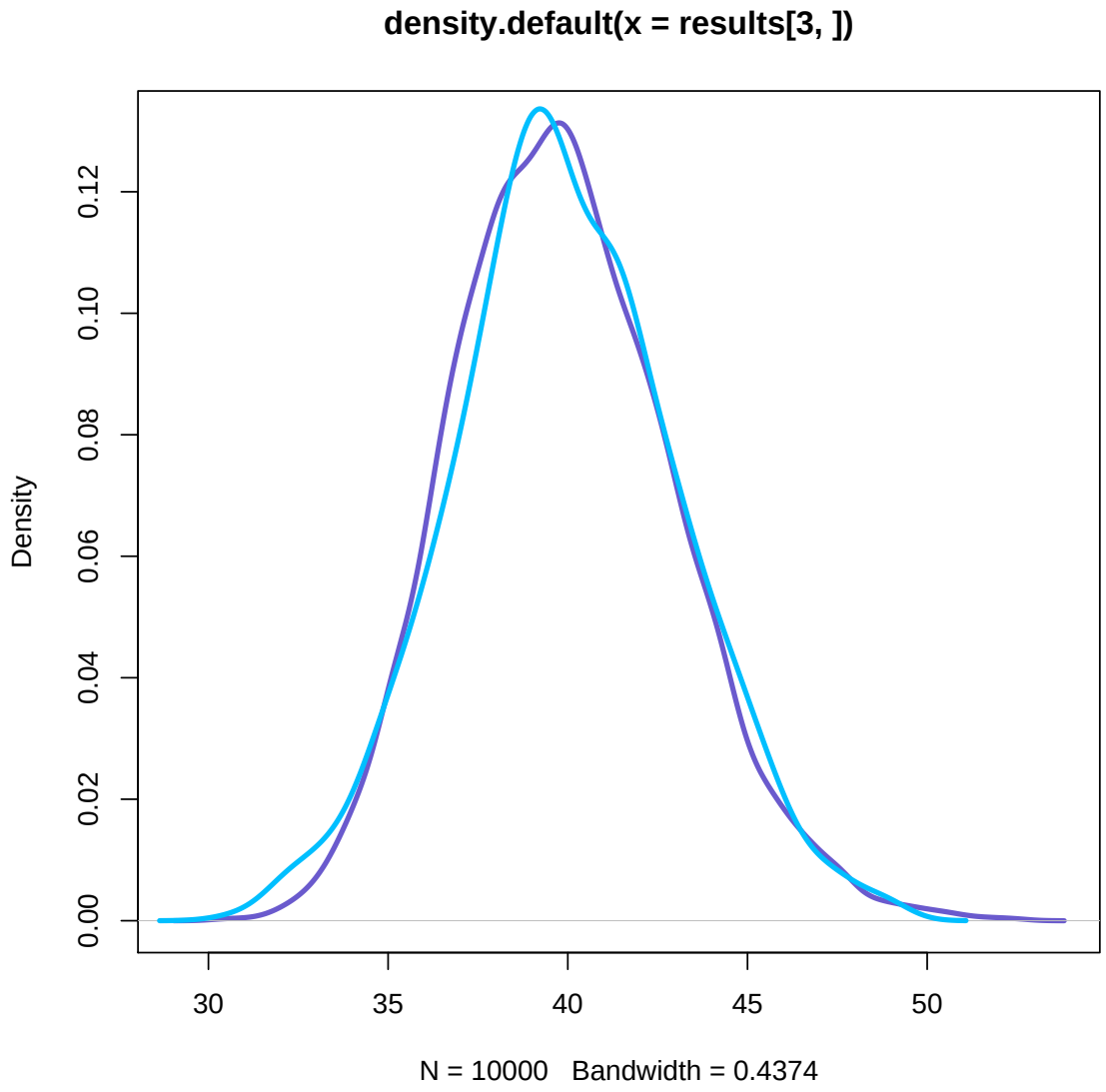
The SE is a good estimator of the sd of the sampling distribution of $\widehat{\beta}$.

Theoretical standard error of $\widehat{\beta}$ in this simulation, i.e. based on $\sigma^2[X'X]^{-1}$.

```
X <- cbind(1, x1)

varcov <- sigma2 * solve(t(X)%*%X)

sqrt(varcov[2,2])

## [1] 0.07566259
```

Plot the distribution of the t-statistics. How close is the theoretical distribution of t to normal? Superimpose a normal curve on the plot (with same mean as t).

```
# Distribution of the t-statistic.

plot(density(results[3,]), lwd=3, col="slateblue")
lines(density(rnorm(1000, mean=mean(results[3,]), sd=sd(results[3,]))),
              lwd=3, col="deepskyblue")
```



7. **(From the 2015 midterm) (The following is based on a real example, but some details are altered for purposes of the question). In November 1993, the state of Pennsylvania conducted elections for its state legislature. The result in the Senate election in the 2nd district (based in Philadelphia) was challenged in court. There, the Democratic candidate won 19,127 of the votes cast by voting machines on election day, while the Republican won 19,691 votes cast by voting machines, giving the Republican a lead of 564 votes. However, the Democrat won 1,396 absentee ballots, while the Republican won just 371 absentee ballots, more than offsetting the Republican lead based on the votes recorded by machines on election day. The Republican candidate sued, claiming that many of the absentee ballots were fraudulent. The judge in the case solicited analysis from an expert, who examined the relationship between absentee vote margins and machine vote**

margins in 21 previous Pennsylvania Senate elections in several districts in the Philadelphia area over the preceding decade. The analysis yielded a data set with the summary statistics depicted in Table 1. The disputed election is not included in the data set.

Table 1: Summary statistics

Variable	Mean	Min	Max
Democratic Margin in Machine Balloting (percentage point lead among two-party votes)	40.68	-13.16	89.32
Democratic Margin in Absentee Ballots (percentage point lead among two-party votes)	29.06	-34.67	72.97

The following characteristics of the data may be useful in answering the questions. Here, X is a 21×2 matrix with 1's in the first column; the second column gives Democratic Margin in Machine Balloting in each election, as defined in Table 1. Also, Y is a 21×1 column vector giving Democratic Margin in Absentee Ballots. We have:

$$X'X = \begin{pmatrix} 21 & 854.21 \\ 854.21 & 53530.38 \end{pmatrix}, \tag{22}$$

$$X'Y = \begin{pmatrix} 610.36 \\ 40597.39 \end{pmatrix}, \tag{23}$$

and

$$Y'Y = 35141.94. \tag{24}$$

(a) **First, use $X'X$ and $X'Y$ to verify the means reported in Table 1. Explain your strategy.**

The (2,1) (and 1,2) element of $X'X$ is 854.21. This gives the sum of Democratic Margin in Machine Balloting (because this is the second column of X pre-multiplied by a row vector of 1s). Thus, the average Democratic Margin in Machine Balloting is $854.21/21 \doteq 40.68$. Similarly, the 1,1 element of $X'Y$, 610.36, is just the sum of the Ys, so the average is $610.36/21 \doteq 29.06$.

(b) **To conduct his analysis, the expert assumed the OLS regression model $Y_i = X_i\gamma + \epsilon_i$ with the usual assumptions. Here, $\gamma = [\alpha \ \beta]'$. What are the usual assumptions? Provide a substantive interpretation for each assumption in this context (i.e., say what they mean, given the details of the application).**

- $Y_i = X_i\gamma + \epsilon_i$.
  This assumption concerns the data-generating process: it says that this equation governs how Y_i is produced as a function of X_i and ϵ_i. We'll say

more about this in item (h), but the equation must here represent a response schedule: if we were to intervene to change Democratic Margin in Machine Balloting by one unit, we would observe an expected change in Democratic Margin in Absentee Balloting of β (the second element of γ). Nothing else influences Y_i beyond X_i and ϵ_i, and the joint distribution of X_i and ϵ_i has the properties described in the next two assumptions.

- $X_i \perp\!\!\!\perp \epsilon_i$ for all i.
  In words, X and ϵ are independent: the value of X gives no information about the value of the random vector ϵ. In particular, shocks to the machine ballot margin are not associated with other shocks that influence absentee margins.
- The ϵ_i are i.i.d. with $E(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$.
  In words, a random shock to the value of Y_i, the Democratic Margin in Absentee Balloting, in one election in one district is independent of the value of the shock in another election or in another district in the same election. Moreover, all these shocks are drawn from the same distribution, which has constant variance σ^2 and a mean of zero.
  See item (i) below for discussion of the plausibility of these assumptions in this context.

(c) **What are the ordinary least squares estimates of α and β? (Do this "by hand" using R or a calculator—do not use matrix manipulations in R such as `solve`. Show your work!). What assumptions of the model do you need to produce these estimates?**

To find the ordinary least squares estimates of α and β, we calculate the coefficient vector using matrix algebra. The coefficient vector is defined as $(X'X)^{-1}X'Y$:

$$
\begin{aligned}
(X'X)^{-1} &= \frac{\text{adj(X'X)}}{\det(\text{X'X})} \\
&= \frac{1}{(21 * 53539.38 - 854.21 * 854.21)} \begin{bmatrix} 53530.38 & -854.21 \\ -854.21 & 21 \end{bmatrix} \\
&= \frac{1}{394463.26} \begin{bmatrix} 53530.38 & -854.21 \\ -854.21 & 21 \end{bmatrix} \\
(X'X)^{-1}X'Y &= \frac{1}{394463.26} \begin{bmatrix} 53530.38 & -854.21 \\ -854.21 & 21 \end{bmatrix} \begin{bmatrix} 610.36 \\ 40597.39 \end{bmatrix} \\
&= \frac{1}{394463.26} \begin{bmatrix} 53530.38 * 610.36 + (-854.21) * 40597.39 \\ (-854.21) * 610.36 + 21 * 40597.39 \end{bmatrix} \\
&= \frac{1}{394463.26} \begin{bmatrix} -2005893.8 \\ 331169.57 \end{bmatrix} \\
&= \begin{bmatrix} -5.0851219 \\ 0.83954479 \end{bmatrix}
\end{aligned}
$$

The estimated intercept, $\hat{\alpha}$, is the (1,1) item of the coefficient vector defined by

$(X'X)^{-1}X'Y$. Thus, $\hat{\alpha} \doteq -5.09$. The estimate for the slope coefficient, $\hat{\beta}$ is the (2,1) item of the coefficient vector. Thus, $\hat{\beta} \doteq 0.84$.

(d) **Provide an interpretation of the estimated intercept $\hat{\alpha}$.**

In the problem we are trying to explain the percentage point lead in absentee ballots of the democratic candidate over the republican with reference to this same margin in ballots recorded by voting machines. The intercept is the baseline difference between the democratic and republican support in absentee ballots; here, it is -5.09, suggesting that when the machine vote Democratic margin is zero (the parties are tied), Democrats fall behind on absentee ballots by 5.09 percentage points. In other words, Democrats are at a relative disadvantage in terms of absentee ballots when 21 elections were studied. This is potentially relevant in the suspect election because the parties are close to tied in machine balloting.

(e) **What are the estimated standard errors of the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$? (Show your work!). What assumptions of the model do you need to justify these standard errors?**

Denote the vector of coefficient estimates by $\hat{\gamma} = (\hat{\alpha}\ \ \hat{\beta})'$. To find the estimated standard errors of the estimates we use the fact that $\widehat{\text{cov}}(\hat{\gamma}|X) = \hat{\sigma}^2(X'X)^{-1}$ as proven by Freedman (2009: 46). The variances for each coefficient will be found on the diagonal of this 2x2 matrix (the matrix is 2x2 because we have estimated two coefficients, $\hat{\alpha}$ and $\hat{\beta}$, and each has its own estimated variance on the diagonal and an estimated covariance on the off-diagonal elements). The estimated standard errors of the estimators are the square root of the diagonal elements of $\widehat{\text{cov}}(\hat{\gamma}|X)$.

Now the problem is to find $\hat{\sigma}^2$, the variance of the regression. One way to estimate the sum of squared errors is by using the sum of squared residuals from the regression, adjusting for the sample size, n and the number of parameters p: $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-p}e'e$, where e denotes the observed vector of residuals. Since we don't have the whole dataset, we don't have the residual vector. A little more algebra is in order. Rearranging the equation for $\hat{\sigma}^2$ we see that:

$$\begin{aligned}(n-p)\hat{\sigma}^2 &= e'e \\ &= (Y - X\hat{\gamma})'(Y - X\hat{\gamma}) \\ &= Y'Y - Y'X\hat{\gamma} - (X\hat{\gamma})'Y + (X\hat{\gamma})'(X\hat{\gamma}) \\ &= Y'Y - (X'Y)'\hat{\gamma} - \hat{\gamma}'(X'Y) + \hat{\gamma}'X'(X\hat{\gamma}).\end{aligned}$$

In the second step we use the fact that $e = Y - X\hat{\gamma}$. We distribute terms in the third step and then in the fourth step, we use the properties of matrix transposes, e.g., $(X\hat{\gamma})' = \hat{\gamma}'X'$.

Now, we are given

$$X'X = \begin{pmatrix} 21 & 854.21 \\ 854.21 & 53530.38 \end{pmatrix}, \tag{25}$$

$$X'Y = \begin{pmatrix} 610.36 \\ 40597.39 \end{pmatrix}, \tag{26}$$

and

$$Y'Y = 35141.94. \tag{27}$$

Next, from (a), we have

$$\hat{\gamma} \doteq \begin{bmatrix} -5.09 \\ 0.84 \end{bmatrix}. \tag{28}$$

Thus,

$$\begin{aligned}
(n-p)\hat{\sigma}^2 &= Y'Y - (X'Y)'\hat{\gamma} - \hat{\gamma}'(X'Y) + \hat{\gamma}'X'(X\hat{\gamma}) \\
&= 35141.94 - \begin{bmatrix} 610.36 & 40597.39 \end{bmatrix} \begin{bmatrix} -5.09 \\ 0.84 \end{bmatrix} - \begin{bmatrix} -5.09 & 0.84 \end{bmatrix} \begin{bmatrix} 610.36 \\ 40597.39 \end{bmatrix} \\
&\quad + \begin{bmatrix} -5.09 & 0.84 \end{bmatrix} \begin{bmatrix} 21 & 854.21 \\ 854.21 & 53530.38 \end{bmatrix} \begin{bmatrix} -5.09 \\ 0.84 \end{bmatrix} \\
&= 35141.94 - [30979.57] - [30979.57] + [30979.57] \\
&= 4162.367
\end{aligned}$$

Since the analysis was carried out with 21 elections, and two parameters were estimated (an intercept and a regression coefficient), $n - p = 21 - 2 = 19$. So we divide 4162.367 by 19 to get our estimate of the sum of squared errors for the regression: $\hat{\sigma}^2 \doteq 219.072$

Then,

$$\begin{aligned}
\widehat{\text{cov}}(\hat{\gamma}|X) &= \hat{\sigma}^2(X'X)^{-1} \\
&= 219.072 \frac{1}{394463.26} \begin{bmatrix} 53530.38 & -854.21 \\ -854.21 & 21 \end{bmatrix} \\
&= \begin{bmatrix} 29.7290 & -0.4744 \\ -0.4744 & 0.0117 \end{bmatrix}.
\end{aligned}$$

The estimated standard error for the estimate of the intercept, $\widehat{\text{SE}}_{\hat{\alpha}}$ is the square root of the (1,1) element of the estimated covariance matrix: $\widehat{\text{SE}}_{\hat{\alpha}} = \sqrt{29.7290} \doteq 5.45$. Similarly, the estimated standard error for the estimated slope coefficient, $\widehat{\text{SE}}_{\hat{\beta}}$ is the square-root of the (2,2) element of the estimated covariance matrix: $\widehat{\text{SE}}_{\hat{\beta}} = \sqrt{0.01117} \doteq 0.108$.

To generate these standard errors we assume that the model that generated the data is $Y = X\gamma + \epsilon$, where $\gamma = (\alpha \ \beta)'$ is the vector of regression coefficients (parameters) and ϵ is a vector of random errors; that the vector of random errors is independent of X; that expected value of the random errors is zero, $\text{E}(\epsilon_\text{i}) = 0$; and most critically, that the random errors are independent and identically distributed with variance $\text{var}(\epsilon) = \sigma^2$. Substantively, the assumption that errors are i.i.d. means that shocks to Democratic margin in one district in Pennsylvania are uncorrelated with shocks to Democratic margin in another district in the state, in a given election; shocks are uncorrelated over time with a district; and so forth. The plausibility of these assumptions is discussed below.

(f) **Conduct a t-test to assess whether the estimate $\hat{\beta}$ is statistically significant. (Here, calculate the t statistic "by hand" using your previous results, then refer the result to the appropriate t distribution).**

The independent variable is the margin in machine balloting. The coefficient of about 0.84 suggests that for every percentage point increase in machine ballot margins, we can expect the absentee margin to rise by 0.84 percentage points. We can conduct a t-test to assess the likelihood of whether, under conditions where the sampling distribution of the coefficient vector $\hat{\gamma}$ is normal, the coefficient we observed is statistically distinct from zero. The t-test is calculated by dividing the coefficient estimate $\hat{\beta}$ by its standard error: $t_{\hat{\beta}} = 0.84/0.108 \doteq 7.77$. For a two-sided t test on 19 degrees of freedom at the 0.05 level, the critical value is 2.09. Thus, 7.77 is way above cutoff for statistical significance.

(g) **Given the model and your estimates, what is the predicted Democratic margin in absentee ballots in the suspect election? What is the standard error of this prediction? Compare the predicted margin to the observed margin. Does it seem likely that something fishy is going on?**

In the purportedly fraudulent election, the Democrats got $19,127$ of the $38,818$ machine votes giving them 49.27 percent of the votes. The margin is $49.27 - (100 - 49.27) = 49.27 - 50.73 = -1.46$ percentage points. With this margin at the machine ballots, the expected difference in absentee ballots given previous elections is -5.08+(0.84)*-1.46 = - 6.31 percentage points.

The estimated variance of this prediction is

$$\widehat{\text{Var}}(\hat{\alpha} + \hat{\beta} * -1.46) = \widehat{\text{Var}}(\hat{\alpha}) + \widehat{\text{Var}}(\hat{\beta}) * (-1.46) + 2(-1.46)\widehat{\text{Cov}}(\hat{\alpha}, \hat{\beta}). \quad (29)$$

(Here, we are treating the margin -1.46 in the suspect election as "fixed", i.e., we are conditioning on the X value; random variation is induced by the assumed error term ϵ, which in turn makes $\hat{\alpha}$ and $\hat{\beta}$ random variables).

From above, $\widehat{\text{Var}}(\hat{\alpha}) = 29.729$; $\widehat{\text{Var}}(\hat{\beta}) = 0.0127$; and $\widehat{\text{Cov}}(\hat{\alpha}, \hat{\beta}) = -0.4744$. Thus, $\widehat{\text{Var}}(\hat{\alpha} + \hat{\beta} * -1.46) = 29.729 + 0.0127 * (-1.46) + 2(-1.46)(-0.4744) = 31.097$. Thus, the estimated standard error for the predicted percentage is about 5.58.

Note that the observed margin in absentee ballots was

$$[1396/(1396 + 371) - (1 - 1396/(1396 + 371))] * 100 = 79 - 21 = 58 \quad (30)$$

percentage points. If the model is right, it seems very unlikely that a margin of 58 percentage points—which is a difference of 64.3 from the predicted percentage—would arise due to chance. After all, this difference is $64.3/5.30 = 12.13$ standard errors away from the predicted percentage. This set-up isn't quite standard as a hypothesis test (we are generating the uncertainty estimates around the predicted percentage not the observed percentage), but it seems very possible—if we believe the model—that something fishy was going on.

(h) **In what ways is this problem like or unlike our example of Hooke's Law? Is this a prediction problem or a causal inference problem? Is a**

***response schedule* required to validate the analysis?**

Like for Hooke's Law, it is critical that the regression model (estimated on data for 21 previous elections in the 2nd and nearby districts of Pennsylvania) govern the response of absentee ballot margins to (changes in) machine ballot margins. If not, the results of the regression analysis are not going to give us a reliable guide to what the relationship should be—absent fraud—in the 2nd district in the suspect election. Thus, a *response schedule* critically undergirds the analysis; if the response schedule is not valid, in the sense of telling us the expected response of absentee balloting to the Democratic margin in machine voting in the suspect election, then neither will conclusions about fraud that are based on this analysis.

Unlike Hooke's Law, here nobody is hanging weights on a spring—or intervening to set the Democratic margin in machine ballots and observing what happens to absentee ballots in consequence. This is an observational study, not an experiment. And the expert may not even be thinking of the "expected response" of absentee balloting in terms of hypothetical interventions to change machine balloting: in this sense, this is a prediction problem, where the goal is to predict (actually postdict) absentee ballot margins as a function of machine ballot margins, under the assumption of no fraud, and compare what should happen to what did happen when fraud was suspected. Even for prediction, though, the response schedule is critical, for the reasons described in the previous paragraph. And causal inference sneaks in the back door, too: estimating the regression equation on data for past elections may not give us a reliable guide to what will happen in the current election, if key factors covary with machine balloting and also influence absentee votes. We expand on this discussion in (i).

(i) **Can you think of any reasons to question the validity of the expert's approach?**

We can begin with the plausibility of the OLS modeling assumptions. One set of concerns surrounds the unbiasedness of the estimates given the model, in particular, the assumption that $X_i \perp\!\!\!\perp \epsilon_i$ for all i is debatable. What influences absentee ballot margins of victory? Republican voters tend disproportionately to vote absentee (as the analysis suggests), as do richer and better-educated voters. Shocks to their voting patterns are likely to influence the Democratic margin in machine as well as absentee balloting, however.

Another set of concerns relates to statistical inference. The model assumes that the ϵ_i are i.i.d. with $E(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$. Thus, shocks in one election are independent across districts, and shocks within a district do not persist from election to election. We did not give precise information about which districts are included in which elections (simply that the analyst looks at 21 elections across several districts in the Philadelphia area in the previous decade), but these assumptions seem highly unlikely. If Republican nominees or Democratic labor unions actively target absentee voters in one election, they are likely to do so across all the districts in the Philadelphia area (or the state). One way to think about the resulting non-sphericality of the error terms is that it reduces our effective N: we do not really have 21 independent observations. Thus, the appearance of statistical

power (and our ability to reject null hypotheses with impressive-looking t statistics) may be misleading). Note also that with 21 units and 19 degrees of freedom, the normal approximations we used for the hypothesis tests are suspect, at least if we are appealing to the central limit theorem. (Given that we are modeling vote share, one might claim the underlying error term is normally distributed, but this is unverifiable—though one could look at the empirical distribution of Y).

Finally, as discussed in item (h), a key aspect of this approach concerns the response schedule: does the equation estimated for the 21 elections previous to the suspect election (and across several districts) give us a reliable guide to the result of a "manipulation" of (change in) democratic margin in the current suspect election, absent fraud? Maybe so. But lots of things change across elections: perhaps the Democratic machine in Philadelphia courted absentee voters especially aggressively in this election, which is good politics but not electoral fraud. One should look at the qualitative evidence carefully to see what might have changed in the 2nd district in this election, relative to previous elections in this and nearby districts.

Incidentally, this question is loosely based on a real problem, involving expert testimony by Princeton economist Orley Ashenfelter. See http://www.nytimes.com/1994/04/11/us/probability-experts-may-decide-pennsylvania-vote.html for media coverage. There, the issue involved a special election for U.S. Senate but several of the analytic issues are similar.

8. **(From the 2015 midterm) In your view, which of these statements is closer to the truth?**

   (a) **Regression analysis can demonstrate causation;**

   (b) **Regression analysis assumes causation but can be used to estimate the size of a causal effect—if the assumptions of the regression models are correct.**

   Pick one of these two statements and defend your choice in detail. Does your answer change, depending on whether we are analyzing experimental or observational data?

   This is for you to argue, and our evaluation is based partly on how well you argue your position. Our view is that option (b) is closer to the truth. When making causal inferences from regression analysis, the *response schedule* plays a key role: it represents a theory of how the data were generated, and it tells us how one variable would respond if we intervened to manipulate other variables. For instance, in multiple regression, the response schedule tells us the functional form of the relationship between the treatment, other measured and unmeasured independent variables, and the outcome. For regression models as well as the standard Neyman model, the response schedule also tells us that the outcome for unit i does not depend on the treatment assignment status of units $j \neq i$. These are substantive assumptions about causal process, and they are built into the model—at least when we are using the regression analysis to draw causal inference.

Given the response schedule, we can use regression to estimate the parameters of the model, for example, the size of a causal effect.

This doesn't really change, whether we are analyzing experimental or observational data. To make causal inferences, we need a response schedule that links treatments to outcomes. In many (but not all) experimental analyses, the response schedule is Neyman's potential outcomes model. Here, regression analysis can provide an unbiased estimator of the average treatment effect, e.g. when we regress the outcome on an intercept and the treatment assignment indicator—which gives us the difference of means between the treatment and control groups. One of the virtues of experiments is that the required assumptions about causal process are often weaker. Yet even for experiments, the underlying modeling assumptions—such as non-interference—play a key role. If those assumptions are wrong, the results can't be trusted. Thus, regression analysis does not demonstrate causation, it makes assumptions about causal process that allow us to estimate the size of causal effects—if the causal assumptions are correct.