Problem Set 4: Linear Regression Modeling

Linh Vu (Collab: Brice Laurent)

Due: October 13, 2023

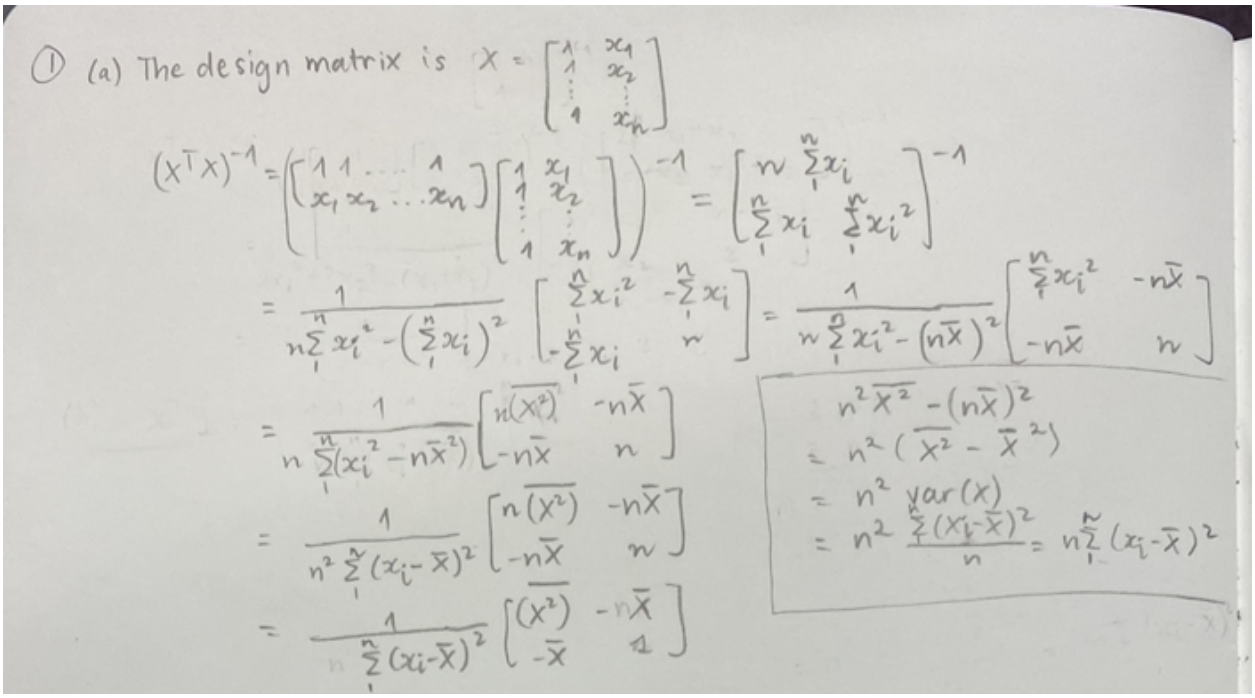
This assignment is **due Friday, October 13 at 11:59pm**, handed in on Gradescope (remember, there are two separate submissions, one for your pdf, and another for you rmd file). Show your work and provide clear, explanations when asked. **Incorporate the relevant R output in this R markdown file**. Only the key output should be displayed for each problem and the relevant parts should be **highlighted** in some way. Make sure that you write-up any interpretation of R-code in your own words (don't just provide the output).

Collaboration policy (for this and all future homeworks): You are encouraged to discuss the problems with other students, but you must write up your solutions yourself and in your own words. Copying someone else's solution, or just making trivial changes is not acceptable.

Problem 1.

Consider a simple linear regression, with an intercept and one predictor.

(a) Write down the design matrix $\mathbf{X}$ and calculate the 2 x 2 matrix $(\mathbf{X}^T\mathbf{X})^{-1}$.



(b) Show that the vector $\hat{\vec{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{Y}$ provides the usual least squares estimates of the intercept and the slope.

(b)

$$\frac{1}{\sum_1^n (x_i-\bar{x})^2}\begin{bmatrix}\overline{x^2} & -\bar{x}\\ -\bar{x} & 1\end{bmatrix}\begin{bmatrix}1 & 1 & \cdots & 1\\ x_1 & x_2 & \cdots & x_n\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}$$

$$= \frac{1}{\sum_1^n (x_i-\bar{x})^2}\begin{bmatrix}\overline{x^2} & -\bar{x}\\ -\bar{x} & 1\end{bmatrix}\begin{bmatrix}\sum_1^n y_i\\ \sum_1^n x_i y_i\end{bmatrix}$$

$$= \frac{1}{\sum_1^n (x_i-\bar{x})^2}\begin{bmatrix}\overline{x^2}\sum_1^n y_i - \bar{x}\sum_1^n x_i y_i\\ -\bar{x}\sum_1^n y_i + \sum_1^n x_i y_i\end{bmatrix}$$

$$\hat{\beta}_1 = \frac{\sum_1^n (x_i-\bar{x})y_i}{\sum_1^n (x_i-\bar{x})^2}$$

$$\hat{\beta}_0 = \frac{1}{\sum_1^n (x_i-\bar{x})^2}\left[\overline{x^2}\sum_1^n y_i - \bar{x}\sum_1^n x_i y_i\right]$$

$$= \frac{1}{\sum_1^n (x_i-\bar{x})^2}\left[\sum_1^n \bar{x}(\bar{x}-x_i)y_i + \sum_1^n \overline{x^2}\, y_i - \bar{x}^2 y_i\right]$$

$$= \bar{x}\cdot\frac{-\sum_1^n (x_i-\bar{x})y_i}{\sum_1^n (x_i-\bar{x})^2} + \frac{\overline{x^2}-\bar{x}^2}{\sum_1^n (x_i-\bar{x})^2}\sum_1^n y_i$$

$$= \bar{x}\cdot -\hat{\beta}_1 + \frac{\mathrm{var}(x)}{\sum_1^n (x_i-\bar{x})^2}\sum_1^n y_i$$

$$= \bar{x}\cdot -\hat{\beta}_1 + \frac{1}{n}\sum_1^n y_i$$

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

(c) Show that, for a simple linear regression, the diagonal elements of the 2 x 2 matrix $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ provide the usual variances of the least squares estimates of the intercept and the slope for a simple linear regression.

We see that the first diagonal entry is equivalent to $\mathrm{Var}(\hat{\beta}_0)$.

(c) Using (a), we know that the 2 diagonal entries of $\sigma^2(X^TX)^{-1}$ are

$$\sigma^2\cdot\frac{1}{(n-1)S_x^2}\cdot\overline{x^2} = \sigma^2\cdot\frac{1}{(n-1)S_x^2}\left[\mathrm{var}(x)+(\bar{x})^2\right] = \sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{(n-1)S_x^2}\right]$$

$$\sigma^2\cdot\frac{1}{(n-1)S_x^2} = \mathrm{var}(\hat{\beta}_1)$$

$$\mathrm{var}(x) = \frac{\sum(x_i-\bar{x})^2}{n}$$

$$(n-1)S_x^2 = \sum(x_i-\bar{x})^2$$

(d) A second predictor is being considered for inclusion in the model (X_2). Under what conditions will its presence in the model have no effect on the estimates of β_0 and β_1?

This happens when X_2 is perfectly uncorrelated with X_1 and has mean 0. None of the variation in Y already explained by X_1 can be explained by X_2, keeping β_1 unchanged, and when X_2 has mean 0, adding X_2 has no effect on the intercept coefficient.

Consider the opposite situation, when the estimates of β_0 and β_1 do change: If we only include X_1 in the model, some of the variation in Y is inappropriately explained by X_1. When we add X_2 to the model (X_2 is correlated with X_1), some of the variation in Y, previously inappropriately explained by X_1, is now appropriately explained by X_2, thereby changing the estimates of β_1. And if X_2 has mean non-zero, the estimate for β_0 needs to be adjusted accordingly when we add X_2 to the model.

Problem 2.

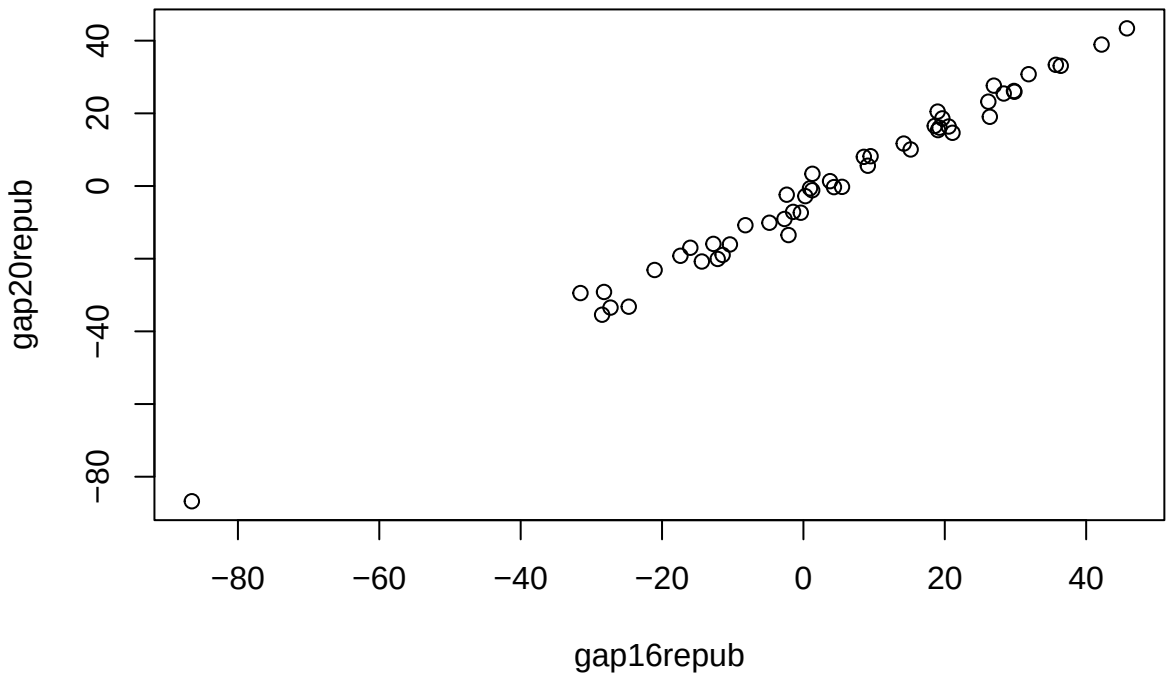
The data set *pres_elections.csv* contains the following variables state by state including DC ($n = 51$)... we will not use all of them:

gap20repub:	the gap in presidential voting from 2020 (Trump - Biden)
gap16repub:	the gap in presidential voting from 2016 (Trump - Clinton)
unemployed:	percent of residents that were unemployed on Nov. 1, 2020
beer:	the average gallons of beer drunk by a state's inhabitants based on a nationwide survey
gmormon:	percent of residents that are Mormon
hispanic:	the percent of a state's inhabitants that consider themselves Hispanic
female:	percent of residents that are women
collegedegree:	percent of adult residents with at least a bachelor's degree
governor:	an indicator for whether the governor of the state was Republican during the election of 2020

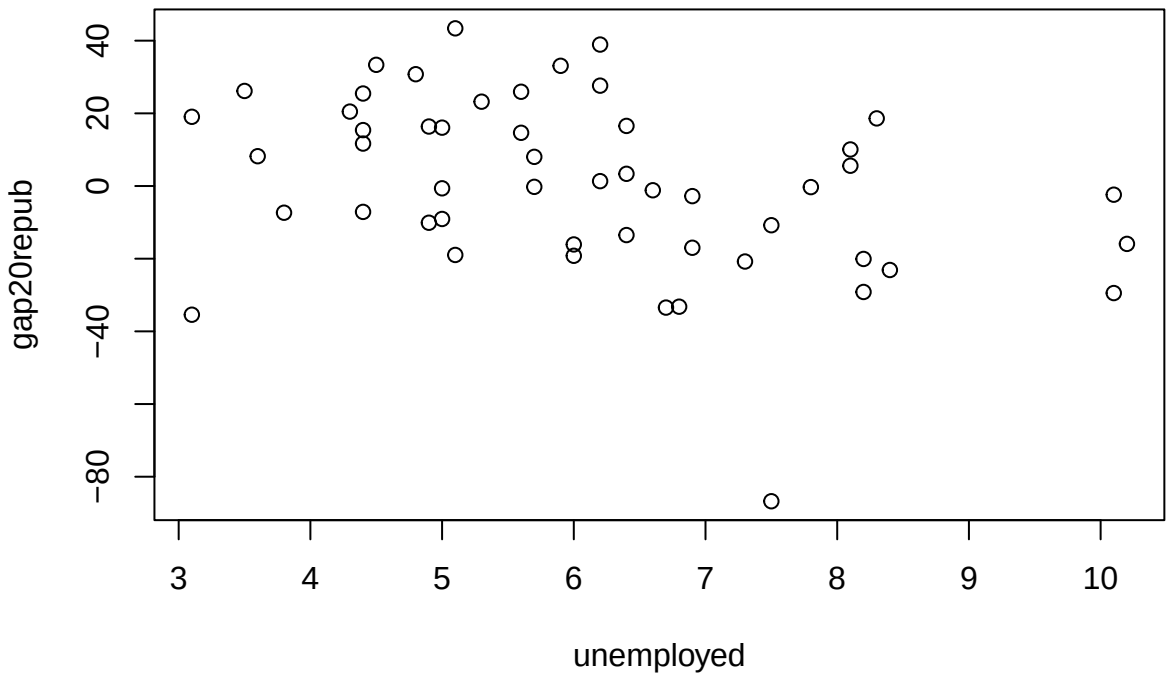
(a) Plot 3 separate scatterplots: $Y =$ `gap20repub` and variables `gap16repub`, `unemployed`, and `beer` as the X variable for each. What do you notice?

In the 1st plot, `gap20repub` and `gap16repub` appear to have a strong positive linear relationship; this makes sense because if a state leaned strongly Republican for 2016, it is likely that they would vote the same in 2020 and vice versa. In the 2nd plot,`gap20repub` and `unemployed` seem to have a very weak, slightly negative correlation. There doesn't seem to be any clear relationship between `gap20repub` and `beer`, per the 3rd plot. There seems to be one outlier – one state that is distinctively more Democratic-leaning than others (apparently it is DC, but this might just be weird data coding).

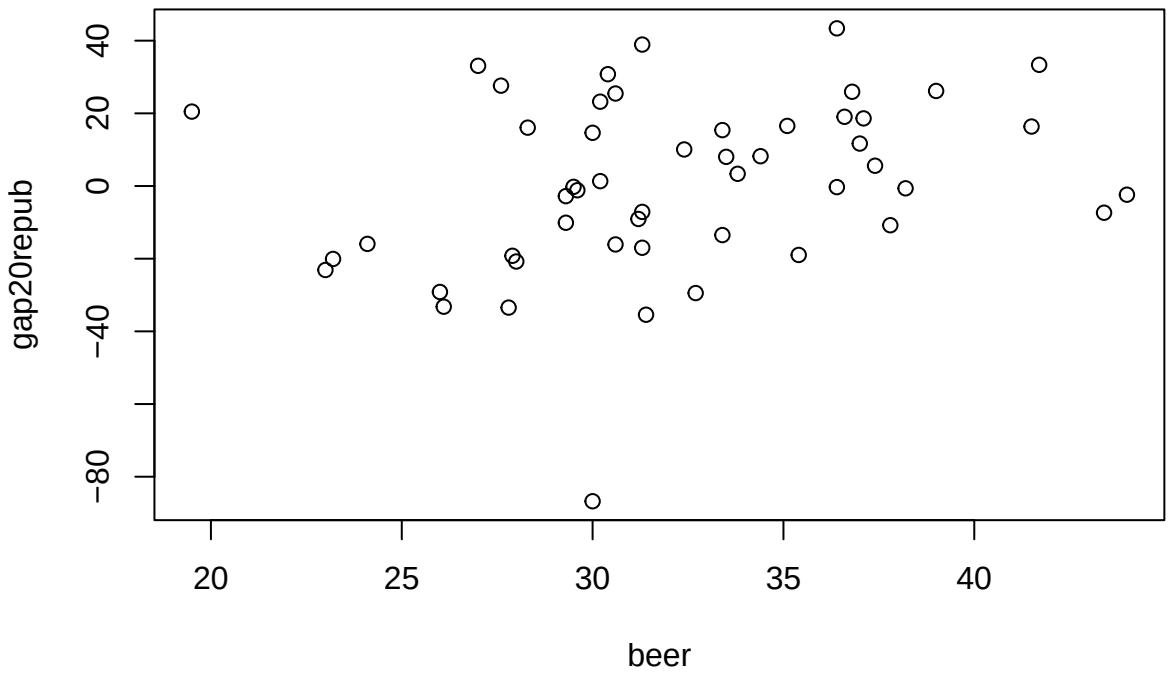
```
pres_elections <- read.csv("data/pres_elections.csv")

plot(gap20repub ~ gap16repub, pres_elections)
```




```
plot(gap20repub ~ unemployed, pres_elections)
```




```
plot(gap20repub ~ beer, pres_elections)
```




```
pres_elections[pres_elections$gap20repub < -60, "state"]
```

```
## [1] "DC"
```

(b) Fit 3 separate regression models to predict `gap20repub`: **Model A** using $X =$ `gap16repub`, **Model B** using $X =$ `unemployed`, and **Model C** using $X =$ `beer`. Include the R summary output and interpret the slope coefficients and their significances.

```
modA <- lm(gap20repub ~ gap16repub, pres_elections)
summary(modA)
```

```
##
## Call:
## lm(formula = gap20repub ~ gap16repub, data = pres_elections)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8842 -2.1394  0.3716  1.8657  5.7775
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.49201    0.40584  -8.604 2.31e-11 ***
## gap16repub   1.00652    0.01704  59.060  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.856 on 49 degrees of freedom
## Multiple R-squared:  0.9861, Adjusted R-squared:  0.9859
## F-statistic:  3488 on 1 and 49 DF,  p-value: < 2.2e-16
```

```
modB <- lm(gap20repub ~ unemployed, pres_elections)
summary(modB)
```

```
##
## Call:
## lm(formula = gap20repub ~ unemployed, data = pres_elections)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.873 -14.122   2.173  15.790  38.936
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   32.761     11.504   2.848  0.00642 **
## unemployed    -5.285      1.817  -2.908  0.00545 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.41 on 49 degrees of freedom
## Multiple R-squared:  0.1472, Adjusted R-squared:  0.1298
## F-statistic: 8.457 on 1 and 49 DF,  p-value: 0.00545
```

```
modC <- lm(gap20repub ~ beer, pres_elections)
summary(modC)
```

```
##
## Call:
## lm(formula = gap20repub ~ beer, data = pres_elections)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.547 -15.114   0.762  15.075  39.491
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -40.1069    20.4467  -1.962   0.0555 .
## beer          1.2635     0.6269   2.015   0.0494 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.32 on 49 degrees of freedom
## Multiple R-squared:  0.07655,    Adjusted R-squared:  0.05771
## F-statistic: 4.062 on 1 and 49 DF,  p-value: 0.04936
```

For model A, the null hypothesis is that the association between `gap20repub` and `gap16repub` is not significant (i.e. the slope coefficient is 0); the alternative hypothesis is that the association is significant. The

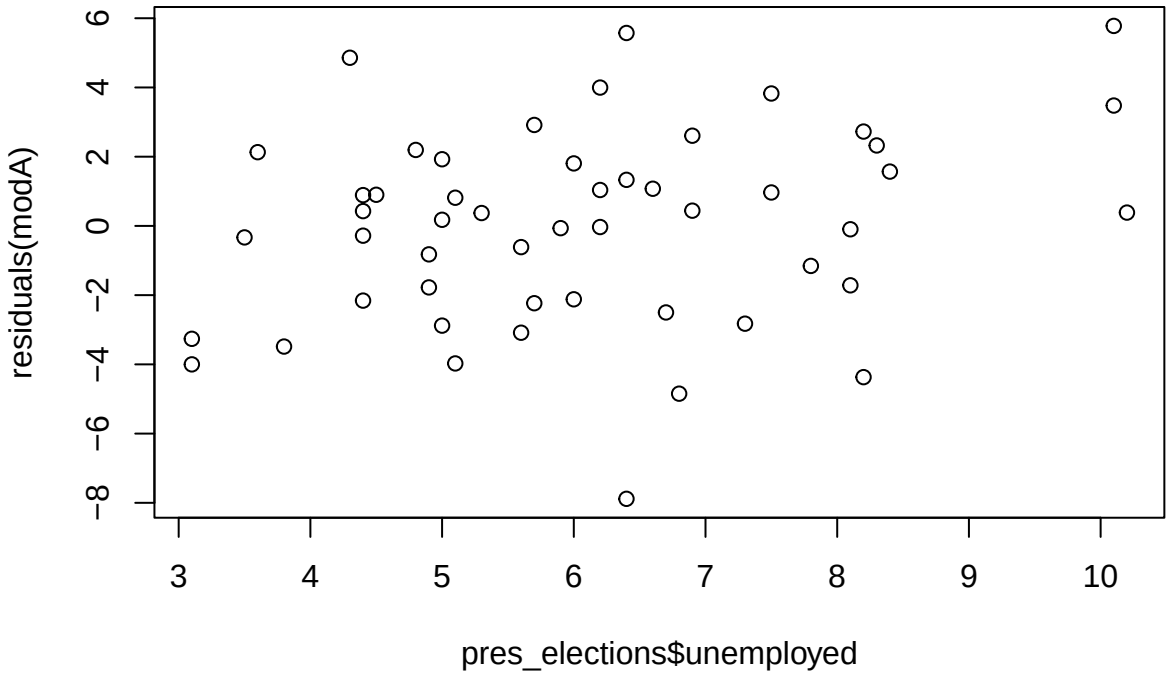
regression table shows that the t statistic=59.0604858 with df=49. The p-value is $3.3524081 \times 10^{-47} < 0.05$, so we reject the null hypothesis and conclude that the association between `gap20repub` and `gap16repub` is significant. The slope coefficient of 1.0065153 means that a 1-unit increase in `gap16repub` is associated with a 1.0065153 unit increase in `gap20repub`.

For model B, the null hypothesis is that the association between `gap20repub` and `unemployed` is not significant (i.e. the slope coefficient is 0); the alternative hypothesis is that the association is significant. The regression table shows that the t statistic=-2.9081282 with df=49. The p-value is $0.0054502 < 0.05$, so we reject the null hypothesis and conclude that the association between `gap20repub` and `unemployed` is significant. The slope coefficient of -5.285098 means that a 1-unit increase in `unemployed` is associated with a -5.285098 unit increase in `gap20repub`.

For model C, the null hypothesis is that the association between `gap20repub` and `beer` is not significant (i.e. the slope coefficient is 0); the alternative hypothesis is that the association is significant. The regression table shows that the t statistic=2.0154809 with df=49. The p-value is $0.0493563 < 0.05$, so we reject the null hypothesis and conclude that the association between `gap20repub` and `beer` is significant. The slope coefficient of 1.2634614 means that a 1-unit increase in `beer` is associated with a 1.2634614 unit increase in `gap20repub`.

(c) Plot the residuals from **Model A** as Y against `unemployed` as X. Fit the regression model for these variables and include the R summary output (call this **Model D**).

```
plot(residuals(modA) ~ pres_elections$unemployed)
```



```
modD <- lm(residuals(modA) ~ pres_elections$unemployed)
summary(modD)
```

```
##
## Call:
## lm(formula = residuals(modA) ~ pres_elections$unemployed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0117 -2.0527  0.4132  1.7064  5.5947
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -2.5053     1.4183  -1.766   0.0836 .
## pres_elections$unemployed   0.4114     0.2240   1.836   0.0724 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.763 on 49 degrees of freedom
## Multiple R-squared:  0.06437,    Adjusted R-squared:  0.04527
## F-statistic: 3.371 on 1 and 49 DF,  p-value: 0.07242
```

(d) Fit a multiple regression model, **Model E** to predict `gap20repub` from `gap16repub` and `unemployed`. Include the R summary output and interpret the slope coefficients and their significances.

```
modE <- lm(gap20repub ~ gap16repub + unemployed, pres_elections)
summary(modE)
```

```
##
## Call:
## lm(formula = gap20repub ~ gap16repub + unemployed, data = pres_elections)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9447 -1.9332  0.4275  1.5777  5.5212
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.58440    1.58563  -4.153 0.000134 ***
## gap16repub   1.02176    0.01819  56.183  < 2e-16 ***
## unemployed   0.49765    0.24718   2.013 0.049706 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.771 on 48 degrees of freedom
## Multiple R-squared:  0.9872, Adjusted R-squared:  0.9867
## F-statistic:  1855 on 2 and 48 DF,  p-value: < 2.2e-16
```

For model E, regarding the predictor `gap16repub`, the null hypothesis is that for states with the same `unemployed` value, the association between `gap20repub` and `gap16repub` is not significant (i.e. the slope coefficient is 0); the alternative hypothesis is that the association is significant. The regression table shows that the t statistic=56.1825689 with df=48. The p-value is $1.8784846 \times 10^{-45} < 0.05$, so we reject the null

hypothesis and conclude that holding the `unemployed` variable constant, the association between `gap20repub` and `gap16repub` is significant. The slope coefficient of 1.0217624 means that if we keep `unemployed` fixed, a 1-unit increase in `gap16repub` is associated with a 1.0217624 unit increase in `gap20repub`.

For model E, regarding the predictor 'unemployed', the null hypothesis is that for states with the same `gap16repub` value, the association between `gap20repub` and `unemployed` is not significant (i.e. the slope coefficient is 0); the alternative hypothesis is that the association is significant. The regression table shows that the t statistic=2.0133286 with df=48. The p-value is 0.0497059 is barely smaller than 0.05, so we reject the null hypothesis and conclude that holding the `gap16repub` variable constant, the association between `gap20repub` and `unemployed` is significant. The slope coefficient of 0.4976513 means that if we keep `gap16repub` fixed, a 1-unit increase in `unemployed` is associated with a 0.4976513 unit increase in `gap20repub`.

(e) Fit a multiple regression model, **Model F**, to predict `gap20repub` from `gap16repub`, `unemployed`, and `beer`. Include the R summary output.

```
modF <- lm(gap20repub ~ gap16repub + unemployed + beer, pres_elections)
summary(modF)
```

```
##
## Call:
## lm(formula = gap20repub ~ gap16repub + unemployed + beer, data = pres_elections)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9512 -1.9409  0.4238  1.5433  5.5778
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.720348   3.097005  -2.170   0.0351 *
## gap16repub   1.021545   0.018860  54.164   <2e-16 ***
## unemployed   0.498767   0.250733   1.989   0.0525 .
## beer         0.004039   0.078727   0.051   0.9593
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.8 on 47 degrees of freedom
## Multiple R-squared:  0.9872, Adjusted R-squared:  0.9864
## F-statistic:  1211 on 3 and 47 DF,  p-value: < 2.2e-16
```

(f) Compare the results from **Model B**, **Model D**, **Model E**, and **Model F** for the coefficient estimates for `unemployed`. Briefly explain the reasons for the results.

For the models, the corresponding coefficient estimates for `unemployed` are -5.285, 0.411, 0.498, 0.499. The coefficient for model B is negative and largest in magnitude: negative because the association between `unemployed` and `gap20repub` is slightly negative (as discussed in plot 2 in part (a)), and large magnitude because this is the only predictor in the model. Model D shows that after `gap16repub` explains the variation in `gap20repub`, the left-over variance can't really be explained by `unemployed` (given the insignificant association between `unemployed` and the residuals of modA). For model E, the coefficient is much smaller than in model B because `gap16repub` is a better predictor for `gap20repub`, shrinking the effect of `unemployed`. Model F has a similar coefficient estimate for `unemployed` to model E: the `beer` doesn't have much predictive power (per our discussion of the 3rd plot in part (a)), so adding it to the model doesn't affect the value of the estimate.

Problem 3.

For the same data set as the previous problem:

(a) Fit a regression model, **Model G**, to predict `gap20repub` from `governor`. Interpret both β coefficient estimates and formally test whether `gap20repub` is associated with `governor`.

```
modG <- lm(gap20repub ~ governor, pres_elections)
summary(modG)
```

```
##
## Call:
## lm(formula = gap20repub ~ governor, data = pres_elections)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.886  -8.746   0.813  13.709  35.254
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -7.933      4.664  -1.701   0.0953 .
## governorRep   16.070      6.409   2.507   0.0155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.85 on 49 degrees of freedom
## Multiple R-squared:  0.1137, Adjusted R-squared:  0.09561
## F-statistic: 6.286 on 1 and 49 DF,  p-value: 0.01553
```

$\hat{\beta_0}$ =-7.933 means that among states without a Republican governor, `gap20repub` is -7.933 units on average. $\hat{\beta_1}$ =16.07 means that states with a Republican governor has a higher `gap20repub` than states without a Republican governor by 16.07 units.

The null hypothesis is that the association between `gap20repub` and `governor` is 0, and the alternative hypothesis is that the association is significant. We conducted a two-sample t test, and the test statistic is 2.507, and df=49. The p-value is 0.016, much smaller than 0.05, so we reject the null hypothesis and conclude that the association between the two variables is significant.

(b) Fit a regression model, **Model H**, to predict `gap20repub` from the predictors `governor`, `gap16repub`, `collegedegree`, and `mormon`. Include the R summary output and interpret the coefficient for `governor`.

```
modH <- lm(gap20repub ~ governor + gap16repub + collegedegree + mormon, pres_elections)
summary(modH)
```

```
##
## Call:
## lm(formula = gap20repub ~ governor + gap16repub + collegedegree +
##     mormon, data = pres_elections)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4394 -1.2923  0.1609  1.3143  4.4212
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.84560    2.70534   4.009 0.000222 ***
## governorRep    0.81029    0.65565   1.236 0.222787
## gap16repub     0.89103    0.02376  37.501  < 2e-16 ***
## collegedegree -0.53147    0.09588  -5.543 1.39e-06 ***
## mormon         0.11214    0.03251   3.449 0.001214 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.177 on 46 degrees of freedom
## Multiple R-squared:  0.9924, Adjusted R-squared:  0.9918
## F-statistic:  1511 on 4 and 46 DF,  p-value: < 2.2e-16
```

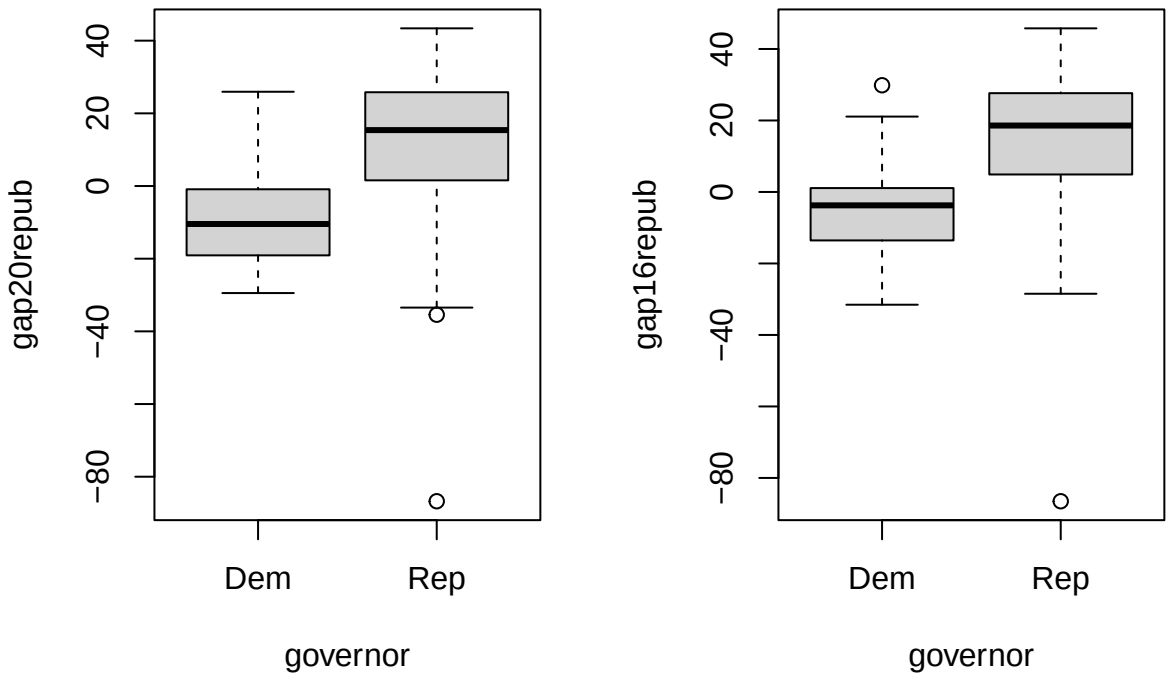
$\hat{\beta}_1$ =0.81 means that holding other variables constant, we expect a state with a Republican governor to have a `gap20repub` 0.81 units higher than a state with a Democrat governor on average.

(c) Compare the coefficient estimate for `governor` in **Model G** and **Model H**. Use the data to explain why there may be any differences or similarities in these estimates.

The coefficient estimate for `governor` is much larger in magnitude in Model G than in Model H; both estimates are positive (which makes sense given the 1st plot). Model G has a larger estimate because the model only includes 1 predictor, while model H includes many other predictors, one of which, `gap16repub`, is strongly associated with both `gap20repub` and `governor` (as shown in plot right). Because much of the explanative power of `governor` is already covered by `gap16repub`, model H has a smaller coefficient estimate for `governor`.

Moreover, according to the covariance matrix below, the covariance between `governorRep` and other variables are negative, meaning that when we include those variables in the model H, the estimate for `governorRep` goes down.

```
par(mfrow=c(1,2))
boxplot(gap20repub ~ governor, pres_elections)
boxplot(gap16repub ~ governor, pres_elections)
```




```
vcov(modH)
```

```
##                (Intercept)  governorRep    gap16repub collegedegree
## (Intercept)    7.318858914  0.005749477 -0.0505702824 -0.2557254691
## governorRep    0.005749477  0.429878607 -0.0043225310 -0.0075755430
## gap16repub    -0.050570282 -0.004322531  0.0005645505  0.0018499614
## collegedegree -0.255725469 -0.007575543  0.0018499614  0.0091923016
## mormon         0.011468031 -0.001954132 -0.0001537016 -0.0004849305
##                       mormon
## (Intercept)    0.0114680311
## governorRep   -0.0019541317
## gap16repub    -0.0001537016
## collegedegree -0.0004849305
## mormon         0.0010570761
```

(d) Use **Model H** to calculate both the 95% confidence and prediction intervals for `gap20repub` in Massachusetts. How do they compare? How does the actual observed `gap20repub` for Massachusetts compare to these intervals? Is this surprising? Why or why not?

```
newdata <- pres_elections[pres_elections$state=="Massachusetts",c("governor", "gap16repub", "collegedeg

predict(modH, newdata, interval="confidence")
```

```
##          fit       lwr       upr
## 22 -32.91901 -34.49581 -31.34221
```

```
predict(modH, newdata, interval="prediction")
```

```
##          fit       lwr       upr
## 22 -32.91901 -37.57619 -28.26183
```

```
pres_elections[pres_elections$state=="Massachusetts", "gap20repub"]
```

```
## [1] -33.46
```

The 95% confidence interval for `gap20repub` in Massachusetts is (-34.496, -31.342), and the 95% prediction interval is (-37.576, -28.262). The point estimate for both intervals is -32.919. The confidence interval is narrower than the prediction interval because the former is a range of possible values for the mean, while the latter is a range of possible values for an unknown value and has more randomness. The observed value of -33.46 falls into the prediction interval, which is as we expected. The observed value also falls into the confidence interval, which is a bit surprising because a certain confidence interval tends to include the true mean but not necessarily a certain observed value.

(e) Perform a formal contrast test based on **Model H** to determine whether the mean `gap20repub` for mythical states with the same predictors as Massachusetts is significantly different than the mean for mythical states with the same predictors as California. Hint: extracting the relevant rows from `model.matrix(modelH)` could be helpful.

```
# get design matrix
X = model.matrix(modH)

# two obs
diff = X[pres_elections$state=="Massachusetts",]-X[pres_elections$state=="California",]

# contrast
estimate = t(modH$coef)%*%diff
std.err = sqrt(t(diff) %*% vcov(modH) %*%diff)
contrast.t = estimate/std.err; contrast.t
```

```
##           [,1]
## [1,] -2.990953
```

```
# p value
2*(1-pt(abs(contrast.t),df=modH$df.residual))
```

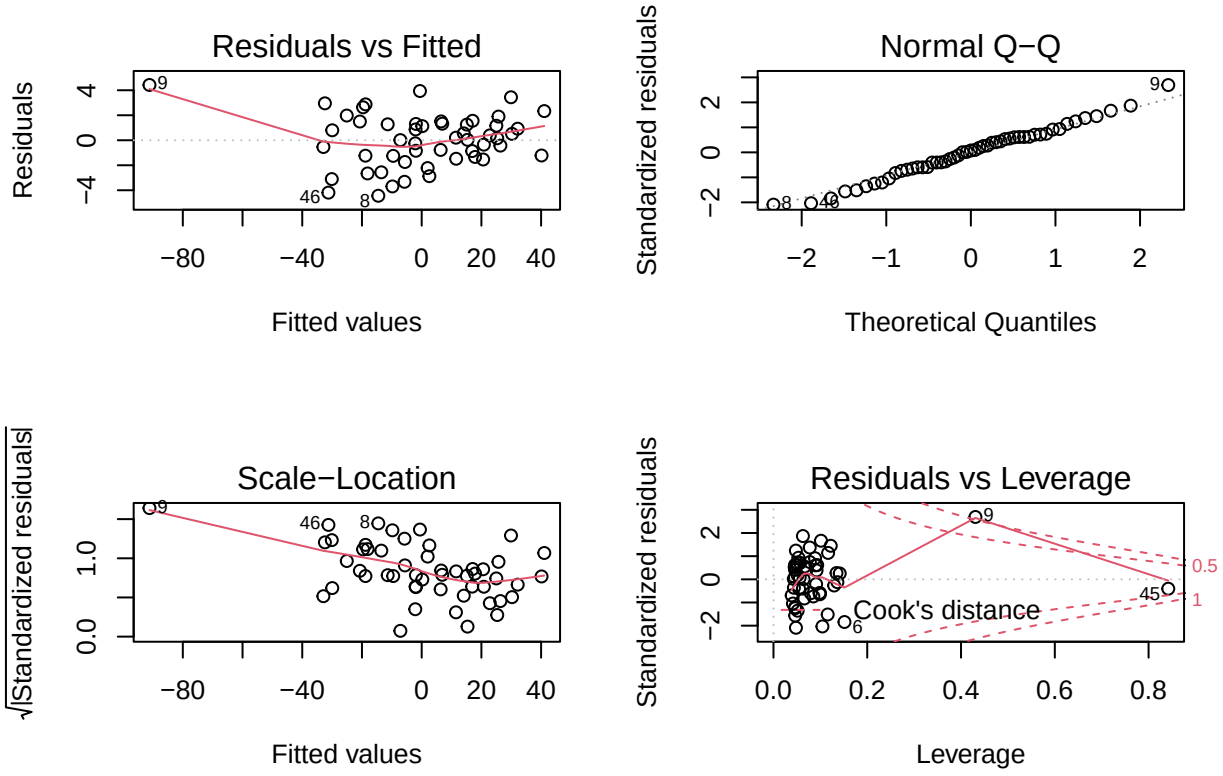
```
##              [,1]
## [1,] 0.004457364
```

We conducted a contrast test, where the null hypothesis is that the mean of `gap20repub` of states similar to Massachusetts is the same as that of states similar to California, and the alternative hypothesis is that the means are different. The test statistic is -2.99, and the p-value is 0.0045. Since the p-value is smaller than 0.05, we reject the null hypothesis and conclude that there is a difference between the means of `gap20repub` of states similar to California and states similar to Massachusetts.

(f) Provide appropriate plots to check the assumptions for **Model H** and comment on whether each assumption is violated and if so, how you would fix the violation (you do not need to implement these fixes). Be specific as to which plot you are using for each assumption.

The normality assumption seems reasonable, as the qqplot of the residuals closely resembles the theoretical quantiles. If we remove the one outlier, the homoscedasticity assumption should be met because it seems like as the fitteed values increase, the residuals vary similarly from the center. And similarly, if we remove the one outlier, the linearity assumption is met because the values for residuals are similar above the center line and below the center line.

```
par(mfrow=c(2,2))
plot(modH)
```



Problem 4.

Perform a simulation study (with 2,000 iterations) where the data are generated from the following model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$ and independent, for $i = 1, 2, ..., n = 50$. Use $\beta_0 = 1$, $\beta_1 = 3$, $\beta_2 = 9$, $\sigma^2 = 10^2$, and sample X_i, Z_i from standard normal distribution where $\text{Cov}(X_i, Z_i) = 0.5$ (but both independent of ε_i).

```
library(mvtnorm)

# params
nsims <- 2000
n <- 50
beta0 <- 1
beta1 <- 3
beta2 <- 9
sigma <- 10
rho <- 0.5
```

(a) For each iteration, fit a correctly specified multiple linear regression model. From these results estimate (i) the coverage probability of the confidence interval for β_1 and (ii) the power of the t-test of $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ for this model.

```
set.seed(139)
cov_prob <- rep(NA, nsims)
pow <- rep(NA, nsims)

for(i in 1:nsims){

  # generate x and z
  Sigma.matrix <- matrix(rep(rho, 4), nrow=2)
  diag(Sigma.matrix) = rep(1,2)
  X <- rmvnorm(n, mean=c(0,0), sigma=Sigma.matrix)
  x <- X[,1]
  z <- X[,2]

  # generate epsilon
  epsilon <- rnorm(n, 0, sigma)

  # generate y
  y <- beta0 + beta1*x + beta2*z + epsilon

  # fit
  mod <- lm(y~x+z)
  cov_prob[i] <- 1*(confint(mod)[2,1] < beta1 & confint(mod)[2,2] > beta1)
  pow[i] <- 1*(summary(mod)$coefficients[2,4] < 0.05)

}
mean(cov_prob)
```

```
## [1] 0.9565
```

```
mean(pow)
```

```
## [1] 0.421
```

(b) For each iteration, fit a misspecified simple linear regression model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. From these results estimate (i) the coverage probability of the confidence interval for β_1 and (ii) the power of the t-test of $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ for this model.

```
set.seed(139)
cov_prob <- rep(NA, nsims)
pow <- rep(NA, nsims)

for(i in 1:nsims){

  # generate x and z
  Sigma.matrix <- matrix(rep(rho, 4), nrow=2)
  diag(Sigma.matrix) = rep(1,2)
  X <- rmvnorm(n, mean=c(0,0), sigma=Sigma.matrix)
  x <- X[,1]
  z <- X[,2]

  # generate epsilon
  epsilon <- rnorm(n, 0, sigma)

  # generate y
  y <- beta0 + beta1*x + beta2*z + epsilon

  # fit
  mod <- lm(y~x)
  cov_prob[i] <- 1*(confint(mod)[2,1] < beta1 & confint(mod)[2,2] > beta1)
  pow[i] <- 1*(summary(mod)$coefficients[2,4] < 0.05)

}
mean(cov_prob)
```

```
## [1] 0.336
```

```
mean(pow)
```

```
## [1] 0.97
```

(c) Interpret the results of this simulation study. What phenomenon does this simulation study highlight? What are the implications to regression modeling in a real data application?

Note: use a confidence level of 0.95 and $\alpha = 0.05$ for all parts.

In part (a), the model has the coverage probability of 0.96, meaning that 96% of the generated confidence intervals include the true value of β_1. The model has the power of 0.42, meaning that it (correctly) rejects the null hypothesis that β_1 is 0 at the rate of 42%.

In part (b), the model has the coverage probability of 0.33, meaning that 33% of the generated confidence intervals include the true value of β_1. The model has the power of 0.97, meaning that it (correctly) rejects the null hypothesis that β_1 is 0 at the rate of 97%.

This simulation study highlights the fact that when we misspecify a model by missing an important confounding variable, the model can give us incorrect estimates and generate CIs with lower coverage probability than we expected, as well as give us a test with inflated power. When we fail to include a confounding variable, the coefficient estimates tend not to be as accurate, so we should be careful to include all confounding variables in regression models.