ST552 Homework 3

Nick Sun

January 29, 2019

Part 1

Say we have the following model:

$$y = X\beta + Z\gamma + \epsilon$$

with X_{nxp} and Z_{nxq} are fixed covariate matrices, β and γ are unknown parameter vectors and $E[\epsilon] = 0$ and $Var(\epsilon) = \sigma^2 I$.

If you don't include Z as a covariate, you fit the following model:

$$y = X\beta + \epsilon$$

(a) Find the expected value of your estimates.

$$\begin{aligned}
E[\hat{\beta}] &= E[(X^T X)^{-1} X^T y] \\
&= E[(X^T X)^{-1} X^T (X\beta + Z\gamma + \epsilon)] \\
&= E[(X^T X)^{-1} X^T X\beta] + (X^T X)^{-1} X^T Z\gamma + (X^T X)^{-1} X^T X^T \epsilon] \\
&= \beta + E[(X^T X)^{-1} X^T Z\gamma] + E[(X^T X)^{-1} X^T \epsilon] \\
&= \beta + E[(X^T X)^{-1} X^T Z\gamma] + (X^T X)^{-1} E[\epsilon] \\
&= \beta + E[(X^T X)^{-1} X^T Z\gamma] \\
&= \beta + (X^T X)^{-1} X^T Z)\gamma
\end{aligned}$$

(b) When is $\hat{\beta}$ unbiased?

It is unbiased in two scenarios:

- Trivially, $\gamma = 0$
- Or when $X^T Z = 0$ i.e. when the rows and columns of Z and X^T are orthogonal.

Part 2

Question 1

A sensible place to being is to investigate wage, education, and experience split amongst our several categorical variables. Note that we have turned region into a single factor variable instead of 4 dummy variables. Using a function like `mosaic::favstats()`, we can get a quick summary of our continuous variables, split by subgroup.

These tables provide us with some interesting information:

- First, the **region** of the worker seems to not matter - wages, education, and experience in each of the four regions seems about the same.
- Second, **years of education** doesn't seem to vary between races, metropolitan residence, or full time/part time workers.
- Third, **years of experience** doesn't seem to vary between races, metropolitan residence, but does seem to vary with part-time/full-time status.
- Fourth, and most notably though, it seems that **wage** in particular is influenced by **race**, **smsa**, and **part time** status.

Table 1: Wage vs. Race

race	min	Q1	median	Q3	max	mean	sd	n
white	50.39	315.805	522.32	795.5875	7716.05	620.9838	468.2589	1844
black	52.23	237.420	398.46	641.0300	2374.15	456.0363	307.5330	156

Table 2: Wage vs. SMSA

smsa	min	Q1	median	Q3	max	mean	sd	n
notsmsa	54.61	260.8025	427.35	664.77	2374.15	497.8030	338.5210	488
smsa	50.39	333.2725	547.47	830.96	7716.05	643.7221	487.4445	1512

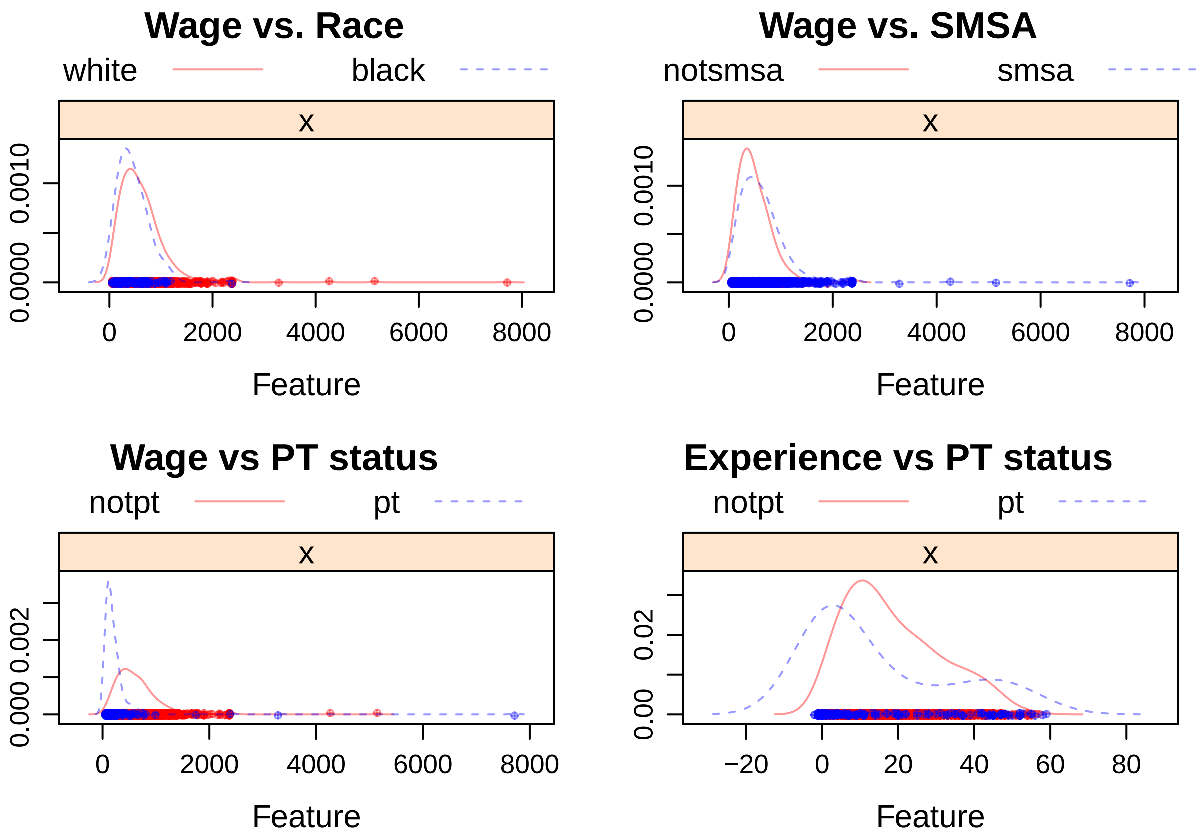
Table 3: Wage vs Part-time status

pt	min	Q1	median	Q3	max	mean	sd	n
notpt	53.83	356.13	557.93	807.22	5144.03	641.7237	422.5631	1815
pt	50.39	96.23	148.15	259.26	7716.05	278.4176	645.2753	185

Table 4: Experience (in years) vs. Part-time status

pt	min	Q1	median	Q3	max	mean	sd	n
notpt	-1	9	16	27	57	18.74325	12.67054	1815
pt	-2	0	5	30	59	15.14595	18.68606	185

Now that we have found some interesting paths of exploration, we can generate density curves of the wage data split up by our categorical variables of interest. We will also generate a density curve of years of experience split up by part-time status since there also seemed to be a relationship there.



These charts tell us a few things: first, the distribution of black wages is centered left of the distribution of white wages. Additionally, the distribution of black workers wages is much less spread out than white wages. Perhaps a symptom of this is that whites have a lot of outliers towards the extremes of wealth while blacks do not. We see a similar story with workers in metropolitan areas; while the distribution is quite similar to works who are not in metropolitan areas but they have several outlier towards to extremes of the distribution.

There is a clear difference in distribution with part-time vs non part-time workers though. As one might expect, part-time workers on average earn far less per week than non part time workers. Lastly, there is also a clear difference in the distribution of years of experience between part time and non part time workers. Non part time workers on average have more years of experience, but interestingly there are more part time workers with 50+ years of experience than non part time workers. A possible example for this might be that non part time workers retire earlier and leave the workforce, or that older folks take part time jobs after retiring from their careers.

Question 2: Faraway 2.7

We are examining the *wafer* dataset which contains 4 categorical explanatory variables and a continuous response variable.

```
##   (Intercept) x1+ x2+ x3+ x4+
## 1           1   0   0   0   0
## 2           1   1   0   0   0
## 3           1   0   1   0   0
## 4           1   1   1   0   0
## 5           1   0   0   1   0
```

Using the model.matrix() function, we can see that `+` is coded as 1 and `-` is coded as 0.

Compute the correlation in the X matrix

```r
cor(model.matrix(wfmodel))
```

```
## Warning in stats::cor(x, y, ...): the standard deviation is zero
```

```
##             (Intercept) x1+ x2+ x3+ x4+
## (Intercept)           1  NA  NA  NA  NA
## x1+                  NA   1   0   0   0
## x2+                  NA   0   1   0   0
## x3+                  NA   0   0   1   0
## x4+                  NA   0   0   0   1
```

We see here that in the rows and columns associated with the Intercept parameter, R reports a value of NA. If we recall the formula for correlation:

$$cor(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

This value will be undefined if either σ_X or σ_Y is 0. The output of model.matrix() shows us that the Intercept variable only ever takes on the number 1, therefore its variance and standard deviation are 0 – it never varies!

Seeing this, the cor() function reports these undefined values as NA.

What difference in resistance is expected when moving from low to the high level of X1?

```r
summary(wfmodel)
```

```
##
## Call:
## lm(formula = resist ~ x1 + x2 + x3 + x4, data = wafer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.381 -17.119   4.825  16.644  33.769
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   236.78      14.77  16.032 5.65e-09 ***
## x1+            25.76      13.21   1.950 0.077085 .
## x2+           -69.89      13.21  -5.291 0.000256 ***
## x3+            43.59      13.21   3.300 0.007083 **
## x4+           -14.49      13.21  -1.097 0.296193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.42 on 11 degrees of freedom
## Multiple R-squared:  0.7996, Adjusted R-squared:  0.7267
## F-statistic: 10.97 on 4 and 11 DF,  p-value: 0.0007815
```

If holding all other variables constant, the average expected resistance increase from flipping x1 from "high" to "low" is 25.76 ohms.

Refit the model without x4 and examine the regression coefficients and standard errors. What stayed the same? What did not?

```
wfmodel2 <- lm(resist ~ x1 + x2 + x3, data = wafer)
wfmodel$coefficients; wfmodel2$coefficients
```

```
## (Intercept)         x1+         x2+         x3+         x4+
##    236.7813     25.7625    -69.8875     43.5875    -14.4875

## (Intercept)         x1+         x2+         x3+
##    229.5375     25.7625    -69.8875     43.5875
```

The parameter estimates for x1, x2, and x3 stayed the same between `wfmodel` and `wfmodel2`. The intercepts however changed, as did the standard error calculations for all of the parameters, the model R^2 values, F-statistic and corresponding p-value, and the calculated $\hat{\sigma}$ and residuals.

Explain how the change in the regression coefficients is related to the correlation matrix of X.

```
cor(model.matrix(wfmodel2))
```

```
## Warning in stats::cor(x, y, ...): the standard deviation is zero

##             (Intercept) x1+ x2+ x3+
## (Intercept)           1  NA  NA  NA
## x1+                  NA   1   0   0
## x2+                  NA   0   1   0
## x3+                  NA   0   0   1
```

From the correlation matrices, we see that the explanatory variables x1, . . ., x4 are not correlated with one another. This would explain why removing one of the variables does not impact the parameter estimates for the remaining variables. However, these explanatory variables must have some correlation with the intercept estimate since the intercept changes as we remove variables.

The `vcov()` function confirms our suspicions. There is a definite relationship between the intercept and the explanatory variables in the model which changes when we remove variables from the model

```
vcov(wfmodel)
```

```
##             (Intercept)           x1+           x2+           x3+
## (Intercept)   218.12520 -8.725008e+01 -8.725008e+01 -8.725008e+01
## x1+           -87.25008  1.745002e+02 -4.843352e-15  9.686705e-15
## x2+           -87.25008 -4.843352e-15  1.745002e+02  4.843352e-15
## x3+           -87.25008  9.686705e-15  4.843352e-15  1.745002e+02
## x4+           -87.25008  9.686705e-15  4.843352e-15  9.686705e-15
##                       x4+
## (Intercept) -8.725008e+01
## x1+          9.686705e-15
## x2+          4.843352e-15
## x3+          9.686705e-15
## x4+          1.745002e+02
```

```r
vcov(wfmodel2)
```

```
##             (Intercept)           x1+           x2+           x3+
## (Intercept)   177.44911 -8.872456e+01 -8.872456e+01 -8.872456e+01
## x1+           -88.72456  1.774491e+02 -4.925202e-15  9.850405e-15
## x2+           -88.72456 -4.925202e-15  1.774491e+02  4.925202e-15
## x3+           -88.72456  9.850405e-15  4.925202e-15  1.774491e+02
```

Question 3

From the *teengamb* dataset from HW2, find the estimate of σ. Interpret this value in context.

```r
data(teengamb)
teenmodel <- lm(gamble ~ sex + status + income + verbal, data = teengamb)
(teensummary <- summary(teenmodel))
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

```r
sigmahat <- teensummary$sigma
```

In the summary() function, $\hat{\sigma}$ is reported as Residual standard error. In the case of the above, $\hat{\sigma} =$ 22.69 pounds. $\hat{\sigma}$ is interpreted as the variance of our residuals, which is to say under our assumptions, the residuals of this model $\epsilon_i \sim N(0, 22.69^2)$.

Find the estimated variance-covariance matrix of the coefficient estimates using matrix algebra

The variance-covariance matrix for $\hat{\beta}$ can be found as:

$$\begin{aligned}
Var[\hat{\beta}] &= Var[(X^TX)^{-1}X^Ty] \\
&= (X^TX)^{-1}X^TVar[y]((X^TX)^{-1}X^T)^T \\
&= (X^TX)^{-1}X^TI_n\sigma^2X(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}$$

Therefore, an estimate for the variance-covariance matrix can be calculated using $\hat{\sigma}$

```
design_matrix <- cbind(rep(1,47),
                       teengamb$sex,
                       teengamb$status,
                       teengamb$income,
                       teengamb$verbal)

xtx <- t(design_matrix) %*% design_matrix
inv_xtx <- solve(xtx)
inv_xtx
```

```
##              [,1]         [,2]          [,3]          [,4]          [,5]
## [1,]  0.574398624 -0.141267359 -0.0046525629 -0.0192062967 -0.0294922334
## [2,] -0.141267359  0.130955021  0.0024738801  0.0047880694 -0.0068775134
## [3,] -0.004652563  0.002473880  0.0001534883  0.0001877384 -0.0006249466
## [4,] -0.019206297  0.004788069  0.0001877384  0.0020421990 -0.0001052899
## [5,] -0.029492233 -0.006877513 -0.0006249466 -0.0001052899  0.0091642662
```

```
(cov_matrix <- (22.69**2) * diag(5) %*% inv_xtx)
```

```
##            [,1]       [,2]        [,3]        [,4]         [,5]
## [1,] 295.721148 -72.729536 -2.39530734 -9.88809489 -15.18366642
## [2,] -72.729536  67.420372  1.27364277  2.46507098  -3.54079217
## [3,]  -2.395307   1.273643  0.07902132  0.09665450  -0.32174507
## [4,]  -9.888095   2.465071  0.09665450  1.05139774  -0.05420707
## [5,] -15.183666  -3.540792 -0.32174507 -0.05420707   4.71809505
```

We can double check this variance-covariance matrix by comparing the variance estimates from the summary of lm() to the diagonal entries in the variance-covariance matrix.

```
(teensummary$coefficients[,2]**2)
```

```
## (Intercept)         sex      status      income      verbal
## 295.7300488  67.4224018   0.0790237   1.0514294   4.7182371
```

```
(diag(cov_matrix))
```

```
## [1] 295.72114758  67.42037245   0.07902132   1.05139774   4.71809505
```

Looks like we calculated our variance-covariance matrix correctly!