STATS 413 PROBLEM SET 3

This problem set is due at **noon ET** on **Sep 30, 2021**. Please upload your solutions to Canvas in two files: a PDF file containing the solutions and a ZIP file containing code that reproduces any computer output in the solutions. You are encouraged to collaborate on problem sets with your classmates, but the final write-up (including any code) **must be your own**.

1. Deviation-from-the-mean regression. Consider a fitting linear model that includes an intercept term. This is equivalent to augmenting the vector of covariates with a constant feature and the vector of regression coefficients with the intercept term:

$$\mathbb{E}\big[\mathbf{y}_i \mid \mathbf{x}_i\big] = \beta_1^* + x_i^T\beta_2^* = \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix}^T \begin{bmatrix} \beta_1^* \\ \beta_2^* \end{bmatrix}.$$

The (augmented) matrix of covariates and vector of OLS coefficients (including the intercept term) has the form

$$\mathbf{X} = \begin{bmatrix} 1_n & \mathbf{X}_2 \end{bmatrix} = \begin{bmatrix} | & — & \mathbf{x}_1^T & — \\ 1_n & & \vdots & \\ | & — & \mathbf{x}_n^T & — \end{bmatrix}, \quad \widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{bmatrix} \begin{matrix} \leftarrow \text{ scalar} \\ \leftarrow \ p-1\text{-vector} \end{matrix},$$

where $1_n \in \mathbf{R}^n$ is the vector of length n whose entries are all one and $\mathbf{X}_2$ is the matrix whose rows are the $\mathbf{x}_i$'s.

(a) Show that the normal equations $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = 0$ are equivalent to

$$\bar{\mathbf{y}} - \widehat{\boldsymbol{\beta}}_1 - \bar{\mathbf{x}}_2^T\widehat{\boldsymbol{\beta}}_2 = 0,$$
$$\frac{1}{n}\mathbf{X}_2^T\mathbf{y} - \bar{\mathbf{x}}_2\widehat{\boldsymbol{\beta}}_1 - \frac{1}{n}\mathbf{X}_2^T\mathbf{X}_2\widehat{\boldsymbol{\beta}}_2 = 0_{p-1}.$$

where $\bar{\mathbf{y}} = \frac{1}{n}\mathbf{y}^T 1_n = \frac{1}{n}\sum_{i=1}^n \mathbf{y}_i$ is the sample mean of the response, $\bar{\mathbf{x}}_2 = \frac{1}{n}\mathbf{X}_2^T 1_n$ is the (vector of) column means of the covariates (excluding the constant covariate), and 0_{p-1} is the vector of length $p-1$ whose entries are all zeros.

Solution: First note the general case:
Given $X^TX\beta = X^Ty$, $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix}$ and $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$.
We can partition into the following:

(1) $X_1^TX_1\beta_1 + X_1^TX_2\beta_2 = X_1^Ty$
(2) $X_2^TX_1\beta_1 + X_2^TX_2\beta_2 = X_2^Ty$

Now solving for $\hat{\beta}_1$ from (1), we get $\hat{\beta}_1 = (X_1^TX_1)^{-1}X_1^T(y - X_2\hat{\beta}_2)$
Now we can fill X_1 with 1_n, and we get the following:

$$\hat{\beta}_1 = (1_n^T1_n)^{-1}1_n^T(y - X_2\hat{\beta}_2) \Rightarrow \hat{\beta}_1 = (\tfrac{1}{n})1_n^T(y - X_2\hat{\beta}_2) \Rightarrow 0 = \bar{y} - \bar{x_2}\hat{\beta}_2 - \hat{\beta}_1$$

and from (2)

$$X_2^T 1_n \hat{\beta}_1 + X_2^T X_2 \hat{\beta}_2 = X_2^T y \Rightarrow 0_{p-1} = X_2^T y - \bar{x_2}(n\hat{\beta_1}) - X_2^T X_2 \hat{\beta}_2$$

(b) Let $\mathbf{H}_1 = \frac{1}{n} 1_n 1_n^T$ be the hat matrix of the constant covariate. Show that the OLS estimator excluding the intercept term is $\widehat{\boldsymbol{\beta}}_2 = (\widetilde{\mathbf{X}}_2^T \widetilde{\mathbf{X}}_2)^{-1} \widetilde{\mathbf{X}}_2^T \widetilde{\mathbf{y}}$, where

$$\widetilde{\mathbf{y}} = (I_n - H_1)\mathbf{y} = \mathbf{y} - 1_n \bar{\mathbf{y}} \quad \widetilde{\mathbf{X}}_2 = (I_n - H_1)\mathbf{X}_2$$

are the *deviations from the sample mean* of the response and non-constant features. Thus including an intercept in the linear model is equivalent to centering the features and outcomes. **Hint:** H_1 is the projector onto $\mathsf{span}\{1_n\}$ (so $I_n - H_1$ is the projector onto the orthocomplement of $\mathsf{span}\{1_n\}$).

Solution: We can use $\hat{\beta_1}$ from (1) in (2) (from the solutions above) to solve this:

$$X_2^T 1_n (1_n^T 1_n)^{-1} 1_n^T (y - X_2 \hat{\beta}_2) + X_2^T X_2 \beta_2 = X_2^T y \Rightarrow X_2^T H_1 y - X_2^T H_1 X_2 \hat{\beta}_2 + X_2^T X_2 \hat{\beta}_2 = X_2^T y \Rightarrow X_2^T (I_n - H_1) X_2 \hat{\beta}_2 = X_2^T (I_n - H_1) y \Rightarrow \hat{\beta}_2 = (\tilde{X_2}^T \tilde{X}_2)^- 1 \tilde{X_2}^T \tilde{y}$$

NOTE: $(I_n - H_1)$ is symmetric and idempotent.

2. Problem 10 in ISLR §3.7. Don't do part (h).
Solution:

Setup

```r
library("ISLR")
```

```
## Warning: package 'ISLR' was built under R version 3.6.3
```

```r
head(Carseats[, c("Sales", "Price", "Urban", "US")])
```

```
##   Sales Price Urban  US
## 1  9.50   120   Yes Yes
## 2 11.22    83   Yes Yes
## 3 10.06    80   Yes Yes
## 4  7.40    97   Yes Yes
## 5  4.15   128   Yes  No
## 6 10.81    72    No Yes
```

Part a

```r
linMod = lm(Sales ~ Price + Urban + US, data= Carseats)
summary(linMod)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

Part b

If price increases by $100 and other predictors are held constant, sales decrease by 5.4459 unit sales. Alternatively, when price increases by $100, the number of carseats sold decrease by 5,445.9.

Whether or not it is in a Urban area, it doesn't affect the sales.

A store in the US sales ~1201 more carseats (in average) than a store that is abroad.

Part c

$Y_{Sales} = 13.04 - 0.05X_{Price} - 0.02X_{UrbanYes} + 1.20X_{USYes}$

Part d

Price and USYes

Part e

```r
linMod2 = lm(Sales ~ Price + US, data= Carseats)
summary(linMod2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

Part f

Based on their R-square values, these two models are not very good. Though, the model from (e) fitting the data slightly better. For both, only 24% change in response explained.

Part g

```r
confint(linMod2)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```