Multiple Regression in practice:

1. [15 Pts.] *Consider the least squares fit of* **y** *to the continuous explanatory variables* **x**, **z**, *and* **w**. *Provided below is the matrix* $(\mathbf{X}^T\mathbf{X})^{-1}$:

$\mathbb{V}(\hat{\beta}) = \sigma^2 (X^TX)^{-1}$

```
 2.56 -0.26 -0.28 -0.07
-0.26  0.03  0.03  0.00
-0.28  0.03  0.04  0.00
-0.07  0.00  0.00  0.01
```

Also provided is the summary of the least squares fit.

$t = \frac{\hat{\beta}_i}{\widehat{SE}(\hat{\beta}_i)}$

$\hat{\beta}_i$ $\widehat{SE}(\hat{\beta}_i)$ t

```
lm(formula = y ~ x + z + w, data = myD)

Coefficients:
             Estimate Std. Error t value
(Intercept)  -3.64797    1.61404  -2.260
x             0.09884    XXXXXX   5.518
z             0.38789    0.19174   2.023
w            -0.21235    0.09320  XXXXXX
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.009 on XXXX degrees of freedom
Multiple R-squared:  0.3373
F-statistic: XXXX on XX and 146 DF
```

$n-(p+1)$ equal

$\hat{\sigma} = \sqrt{\frac{e^Te}{n-(p+1)}} = \sqrt{\frac{ESS}{n-(p+1)}}$

$R^2 = \frac{Reg\,SS}{TSS} = \frac{TSS - ESS}{TSS}$

i.e. $ESS = \left(\frac{1}{R^2} - 1\right) \cdot Reg\,SS$

(a) *What are the five missing values indicated by XXXXX in the above summary?*

$p = 3$ in this case

$F = \frac{Reg\,SS/p}{ESS/n-(p+1)} = \frac{R^2}{1-R^2} \cdot \frac{n-(p+1)}{p}$

(b) *Explain how to test whether the coefficient for* **z** *is zero.*

① null hypothesis (H_0)

② t stat

③ If value lies in the extreme tail of the null distribution, reject H_0

(c) *The standard deviations (normalized by* $n-1$*) for the variables in this study are as follows:*

Variable	*SD*
y	*1.23*
x	*10.14*
z	*0.96*
w	*0.91*

Compute, compare, and comment on the standardized regression coefficients for **x** *and* **w**.

Std coefficient of x $\hat{\beta}_x' = \hat{\beta}_x \cdot \frac{SD(x)}{SD(y)}$

Dummy Variable Regression in practice:

1. Let's imagine that 80 students took a particular course at Berkeley of whom 20 were freshmen, 20 were sophomores, 20 were juniors and 20 were seniors. In R, I have saved the final scores (out of 100) for the 20 freshmen in the vector **g1**, for the 20 sophomores in **g2**, juniors in **g3** and seniors in **g4**. Consider the following output:

mean & SD

```
> mean(g1)
[1] 57.96
> sd(g1)
[1] 3.92
> mean(g2)
[1] 64.13
> sd(g2)
[1] 3.91
> mean(g3)
[1] 67.60
> sd(g3)
[1] 6.92
> mean(g4)
[1] 71.22
> sd(g4)
[1] 5.77
```

cal as $(n-1)$

Question	Total points
Q1	12
Q2	6
Q3	9
Q4	6
Q5	12
Q6	14
Q7	12
	60

Also, for $i = 1, \ldots, 80$, let

- y_i: Final score of the i^{th} student in the class.
- x_{i1}: Takes the value 1 if the i^{th} student is a freshman and 0 otherwise.
- x_{i2}: Takes the value 1 if the i^{th} student is a sophomore and 0 otherwise.
- x_{i3}: Takes the value 1 if the i^{th} student is a junior and 0 otherwise.
- x_{i4}: Takes the value 1 if the i^{th} student is a senior and 0 otherwise.

I fit the linear model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i, i = 1, \ldots, n$$

to this data via R to obtain the following output:

① The mean of Base Case

② $t = \frac{\beta_i}{SE(\beta_i)}$

All dummy!

R actually run $y \sim x_1 + x_2 + x_3$ + intercept (senior has excluded, seen as Base Case)

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
     Min       1Q   Median       3Q      Max
-14.4685  -3.6042  -0.1473   3.1631  12.5674

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     XXXX       1.18   60.24  < 2e-16 ***
x1            -13.26       1.67   -7.93  1.5e-11 ***
x2             -7.09       1.67   -4.24  6.2e-05 ***
x3             -3.62       1.67   -2.16    0.034 *
x4                NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: XXXXX on 76 degrees of freedom
Multiple R-squared:  0.4734,Adjusted R-squared:  XXXXX
F-statistic: 22.77 on 3 and 76 DF,  p-value: 1.279e-10
```

$\sqrt{VIF} \cdot \hat{\sigma}\sqrt{V_{ii}}$

$n-(p+1)$ $p=3$, not 4 !

$= 1 - \left(\frac{n-1}{n-(p+1)} \cdot \frac{ErrSS}{TSS}\right)$

$= 1 - \left(\frac{n-1}{n-(p+1)} \cdot (1-R^2)\right)$

SE (SD of means)

$\sqrt{\hat{\sigma}^2} = \sqrt{\frac{e^Te}{n-(p+1)}} = \sqrt{\frac{ESS}{n-(p+1)}}$

$ESS = \sum_i (\text{Size of group}_i - 1) \times \text{Sample Variance}_i$

$= \underbrace{(20-1)\times 3.92 + (20-1)\times 3.91 + \ldots}_{\text{4 groups! not 3}}$

(b) Fill in the 3 missing values in the R output with proper reasoning. (**6 points**).

(a) Why does the R output above say "`1 not defined because of singularities`"? Give reasons for your answer and suggest a way to fix the problem. (**2 points**)

(explain perfect collinearity)

(c) Explain why the standard error estimates for the coefficients of **x1**, **x2**, and **x3** are all the same. (**3 points**).

$$SE(\beta) = \sqrt{\text{diagonal of } \hat{\sigma}^2 (X^TX)^{-1}}$$

$$X: \begin{array}{cccc} \text{intercept} & x_1 & x_2 & x_3 \end{array}$$
$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & 1 & 0 \\ \vdots & 0 & \vdots & 0 \\ \vdots & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ & 0 & 0 & \vdots \end{pmatrix}$$

$$X^TX = \begin{bmatrix} 80 & 20 & 20 & 20 \\ 20 & 20 & & \\ 20 & & 20 & \\ 20 & & & 20 \end{bmatrix}$$