POL 213, Spring 2024
Problem Set 1 Solution

Yu-Shiuan (Lily) Huang

1 Least Squares Fit

Using the following data, with `prestige` as the response variable (y) and `education` as the explanatory variable (x), compute the intercept, regression coefficient, residual standard error, total sum of squares, residual sum of squares, and R-squared *by hand*. Write out the equations you used to calculate each quantity. Show your work. You may, of course, use R to perform your calculations, but you must SHOW YOUR WORK! You will find Fox Chapter 5.1 to be helpful.

b. regression coefficient B

$$B = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} = \frac{(2.2+1.2+27.6-14.2-13.8-15.6+49.5+82.8+31.1-9.8)}{(4+144+16+4+4+16+25+144+1+4)} = \frac{141}{362} = 0.3895028$$

prestige (Y_i)	education (X_i)	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) * (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
82	86	86-88=-2	82-83.1=-1.1	2.2	4
83	76	76-88=-12	83-83.1=-0.1	1.2	144
90	92	92-88=4	90-83.1=6.9	27.6	16
76	90	90-88=2	76-83.1=-7.1	-14.2	4
90	86	86-88=-2	90-83.1=6.9	-13.8	4
87	84	84-88=-4	87-83.1=3.9	-15.6	16
93	93	93-88=5	93-83.1=9.9	49.5	25
90	100	100-88=12	90-83.1=6.9	82.8	144
52	87	87-88=-1	52-83.1=-31.1	31.1	1
88	86	86-88=-2	88-83.1=4.9	-9.8	4

```r
# load data
df <- read.csv("data/ps1_prestige.csv")

# calculate B
b <- sum((df$education-mean(df$education))*(df$prestige-mean(df$prestige)))/
                sum((df$education-mean(df$education))^2)
```

a. intercept A

$$A = \bar{Y} - B\bar{X}$$

$$A = 83.1 - 0.3895028 * 88 = 48.82376$$

```r
# calculate A
a <- mean(df$prestige)-b*mean(df$education)
```

c. residual standard error S_E

$S_E = \sqrt{\frac{\Sigma {E_i}^2}{n-2}} = \sqrt{\frac{1243.98}{10-2}} = 12.46986$

$E_i^2 = (-0.3209945)^2+(4.5740331)^2+(5.3419890)^2+(-7.8790055)^2+(7.6790055)^2+(5.4580110)^2+(7.9524862)^2 + (2.2259669)^2 + (-30.7104972)^2 + (5.6790055)^2 = 1243.98$

prestige (Y_i)	education (X_i)	$\hat{Y}_i = A + BX_i$	$E_i = Y_i - \hat{Y}_i$
82	86	48.82376+0.3895028*86=82.32099	-0.3209945
83	76	48.82376+0.3895028*76=78.42597	4.5740331
90	92	48.82376+0.3895028*92=84.65801	5.3419890
76	90	48.82376+0.3895028*90=83.87901	-7.8790055
90	86	48.82376+0.3895028*86=82.32099	7.6790055
87	84	48.82376+0.3895028*84=81.54199	5.4580110
93	93	48.82376+0.3895028*93=85.04751	7.9524862
90	100	48.82376+0.3895028*100=87.77403	2.2259669
52	87	48.82376+0.3895028*87=82.71050	-30.7104972
88	86	48.82376+0.3895028*86=82.32099	5.6790055

```
# calculate residual standard error
yhat <- a + b*df$education
ei <- df$prestige - yhat
rse <- sqrt(sum(ei^2)/(10-2))
```

d. total sum of squares (TSS)

$\Sigma {E_i'}^2 = \Sigma(Y_i - \bar{Y})^2 = (1.1)^2 + (-0.1)^2 + (6.9)^2 + (-7.1)^2 + (6.9)^2 + (3.9)^2 + (9.9)^2 + (6.9)^2 + (-31.1)^2 + (4.9)^2 = 1298.9$

```
# calculate tss
tss <- sum((df$prestige - mean(df$prestige))^2)
```

e. residual sum of squares (RSS)

$\Sigma {E_i}^2 = \Sigma(Y_i - \hat{Y})^2 = (-0.3209945)^2 + (4.5740331)^2 + (5.3419890)^2 + (-7.8790055)^2 + (7.6790055)^2 + (5.4580110)^2 + (7.9524862)^2 + (2.2259669)^2 + (-30.7104972)^2 + (5.6790055)^2 = 1243.98$

```
# calculate rss
rss <- sum((df$prestige - yhat)^2)
```

f. regression sum of squares (RegSS)

$RegSS = TSS - RSS = 1298.9 - 1243.98 = 54.92$

g. R-squared

$r^2 = \frac{RegSS}{TSS} = \frac{54.92}{1298.9} = 0.04228193$

2 Single and Multivariate Regression

Analyze the data `Anscombe.txt` available on Canvas. The data measures U.S. State Public School Expenditures. I think it is from 1981. The variables are:

- `education` = per capita education expenditures, dollars
- `income` = per capita income, dollars
- `under18` = proportion under 18 years old, per 1000
- `urban` = proportion urban, per 1000.

Using these data, answer the following questions. A helpful tool for creating scatterplots is the `car` package in R.:

a. Draw a separate scatterplot showing the relationship of the response variable `education` to each explanatory variable.

```
# load data
df2 <- read.delim("data/Anscombe.txt", sep = "")

# a. scatterplots
library(car)
scatterplot(education ~ income, data = df2)
scatterplot(education ~ under18, data = df2)
scatterplot(education ~ urban, data = df2)
```

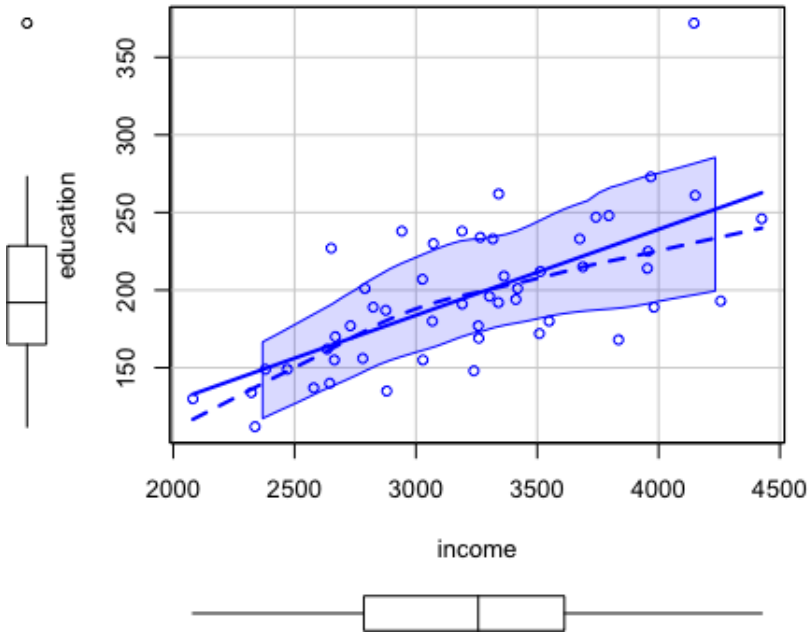

Figure 1: Relationship between `education` and `income`

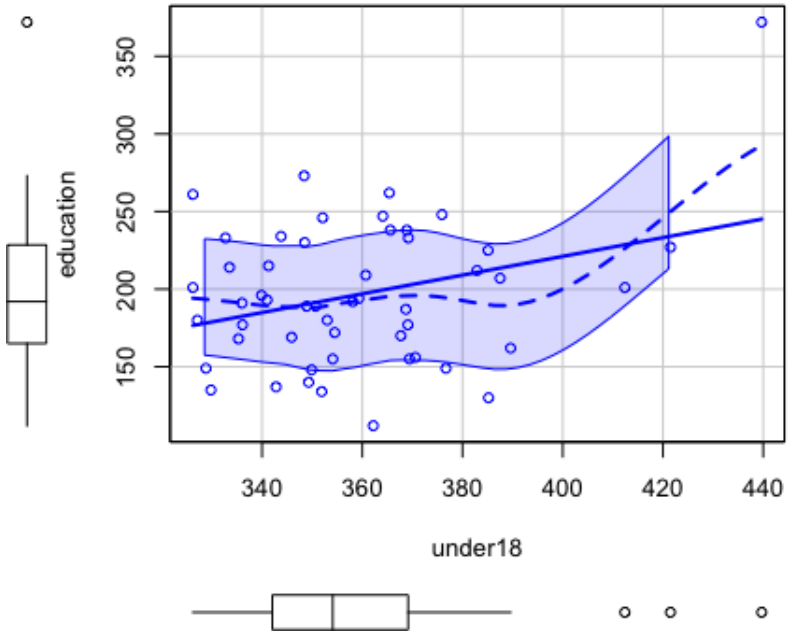

Figure 2: Relationship between `education` and `under18`

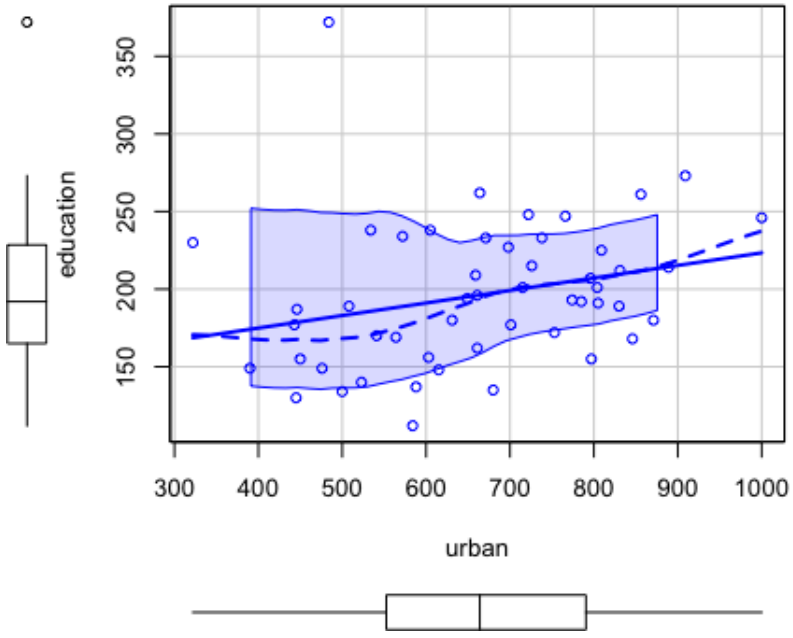

Figure 3: Relationship between `education` and `urban`

b. Compute the simple (bivariate) linear regression of the response on each explanatory variable. Once you have each bivariate regression, report the A, B, SE, and r^2 and substantively interpret each of these quantities.

Table 1 presents the results of a simple linear regression analysis examining the relationship between `education` and each explanatory variable (`income`, `under18`, and `urban`). Model 1 estimates the relationship between `education` and `income`. As shown in Model 1, the expected value of education expenditure is 17.71 when income is 0, and that there is a statistically significant positive relationship between education and income. Specifically, a one-unit increase in per capita income leads to a 0.06 unit increase in education expenditure. The residual standard error, which measures the average vertical distance between the regression line and each observation, provides insight into how closely the estimated regression line aligns with the scatter of points. For Model 1, the residual standard error is 34.94. R^2 is a statistical measure that quantifies the proportion of variance in a response variable that is explained by an explanatory variable in a regression model. In Model 1, 45% of the variance in education expenditure is accounted for by `income`.

Table 1: Relationship between Education and each Explanatory Variable

	Dependent variable:		
	Education		
	(1)	(2)	(3)
Income	0.06***		
	(0.01)		
Under18		0.60*	
		(0.26)	
Urban			0.08
			(0.04)
Constant	17.71	−20.42	142.60***
	(28.87)	(94.66)	(28.82)
Observations	51	51	51
R^2	0.45	0.10	0.07
Adjusted R^2	0.43	0.08	0.05
Residual Std. Error (df = 49)	34.94	44.59	45.27
F Statistic (df = 1; 49)	39.39***	5.26*	3.65
Note:	*p<0.05; **p<0.01; ***p<0.001		

Models 2 and 3 investigate the relationship between `education`, `under18`, and `urban`, respectively. As revealed in the results of Models 2 and 3, when the proportion of individuals under 18 years old and the proportion of urban area are both 0, the expected value of education expenditure is -20.42 and 142.60, respectively. While there is a statistically significant positive relationship between education expenditure and the proportion of individuals under 18 years old, there is no detectable relationship between education expenditure and the proportion of urban area. To be more specific, as the proportion under 18 years old increases by one unit, education expenditure increases by 0.6 units. The residual standard errors in Models 2 and 3 are 44.59 and 45.27, respectively, both indicating a higher average vertical distance between the regression line and each observation than that observed in Model 1. The R^2 values in

Models 2 and 3 are 0.10 and 0.07, respectively, which are notably lower than that in Model 1. Overall, these results suggest that the least squares fit in Model 1 is superior to that in Models 2 and 3.

```
# b. simple linear regression
m1 <- lm(education ~ income, data = df2)
m2 <- lm(education ~ under18, data = df2)
m3 <- lm(education ~ urban, data = df2)

library(stargazer)
stargazer(m1, m2, m3, digits = 2,
          star.cutoffs = c(0.05, 0.01, 0.001))
```

c. Draw the least-squares line on the scatterplot. Is the least-squares line a reasonable summary of the relationship between the two variables?

For reference, please see Figures 1, 2, and 3, with the solid blue line representing the least-squares regression line. Overall, the least-squares line in each model provides a reasonable summary of the relationship between the response variable (`education`) and each explanatory variable. All three figures show a positive relationship between `education` and the respective explanatory variables, consistent with the estimates in Table 1. Moreover, we can observe that the average vertical distance between the regression line and each observation is smaller in Figure 1, but larger in both Figures 2 and 3, which corresponds to our analysis in Table 1.

d. Then run a multiple regression of the response variable, `education`, on all the explanatory variables at once. Does your substantive interpretation of the relationship change? Why or why not?

Model 4 examines the relationship between `education` and all of the explanatory variables simultaneously using multiple regression analysis. The results show some differences compared to Models 1-3. Notably, the coefficient of `urban` that was positive but not statistically significant in Model 3 is now significantly negative in Model 4. Specifically, for each one-unit increase in the proportion of urban area, education expenditure decreases by 0.11 units, holding constant the other explanatory variables. The lack of a significant effect in Model 3 could have been due to the omission of other relevant variables that influence the relationship between `education` and `urban`. Moreover, the least-squares fit improves markedly in Model 4: the residual standard error decreases to 26.96 and the R^2 increases to 0.69. These results suggest that the three explanatory variables jointly explain a substantial portion of the variance in `education`.

```
# d. multiple linear regression
m4 <- lm(education ~ income + under18 + urban, data = df2)
stargazer(m1, m2, m3, m4, digits = 2,
          star.cutoffs = c(0.05, 0.01, 0.001))
```

Table 2: Determinants of Education Expenditure

	Dependent variable:			
	Education			
	(1)	(2)	(3)	(4)
Income	0.06***			0.08***
	(0.01)			(0.01)
Under18		0.60*		0.82***
		(0.26)		(0.16)
Urban			0.08	−0.11**
			(0.04)	(0.03)
Constant	17.71	−20.42	142.60***	−286.84***
	(28.87)	(94.66)	(28.82)	(64.92)
Observations	51	51	51	51
R^2	0.45	0.10	0.07	0.69
Adjusted R^2	0.43	0.08	0.05	0.67
Residual Std. Error	34.94 (df = 49)	44.59 (df = 49)	45.27 (df = 49)	26.69 (df = 47)
F Statistic	39.39*** (df = 1; 49)	5.26* (df = 1; 49)	3.65 (df = 1; 49)	34.81*** (df = 3; 47)
Note:				*p<0.05; **p<0.01; ***p<0.001

e. Finally, devise a sensible hypothesis about one of the explanatory variables and test it with your data. Explain each step of your process.

The alternative hypothesis I want to test is whether a state's education expenditure increases as the proportion of individuals under 18 years old increases. To investigate this, I constructed a model that includes two variables of interest (`education` and `under18`) and two control variables (`income` and `urban`). Next, I ran a multiple linear regression to determine if the coefficient of `under18` is statistically significant at any conventional level. The results, as shown in Model 4 in Table 1, indicate that the coefficient of `under18` is positive (0.82) with a standard error of 0.16. By conducting a t-test, I found that the t-statistic is 5.125 ($t = \frac{B_{under18} - \beta_{under18}}{SE(B_{under18})} = \frac{0.82-0}{0.16} = 5.125$). Since the t-statistic is greater than the critical value ($t_{\frac{0.05}{2}, df=51-2} = 2.01$) at a significance level of $\alpha = 0.05$, we can reject the null hypothesis in favor of our alternative hypothesis.

3 Properties of the Least Squares Estimator

Demonstrate the features of the least squares estimator. Make sure to explain every step in your derivations, justifying why you can move from one equality to the next. Parts (b.) and (c.) are challenge problems, only required of methodology students.

Here are some reviews before starting the below problems:

- The assumed model: $Y_i = \alpha + \beta X_i + \varepsilon_i$
- Linearity assumption: $E(\varepsilon_i) = 0$
- Normality assumption: $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$
- Independence assumption: $Cov(\varepsilon_i, \varepsilon_j) = 0, for i \neq j$
- The expectation of linear combination: $E(a + bY) = a + bE(Y)$
- The variance of linear combination:
  - $Var(a + bY) = b^2 Var(Y)$
  - $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
- For the below demonstration, the X_i are assumed to be fixed, not random.
- $B = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} = \frac{\Sigma(X_i - \bar{X})Y_i - \Sigma(X_i - \bar{X})\bar{Y}}{\Sigma(X_i - \bar{X})^2} = \frac{\Sigma(X_i - \bar{X})Y_i - \bar{Y}\Sigma(X_i - \bar{X})}{\Sigma(X_i - \bar{X})^2} = \frac{\Sigma(X_i - \bar{X})Y_i}{\Sigma(X_i - \bar{X})^2}$
- $\Sigma(X_i - \bar{X})^2 = \Sigma(X_i - \bar{X})(X_i - \bar{X}) = \Sigma X_i(X_i - \bar{X}) - \bar{X}\Sigma(X_i - \bar{X}) = \Sigma X_i(X_i - \bar{X})$

a. Demonstrate the unbiasedness of the least-squares estimators B for β in simple regression. Do this by expressing the least-squares slope B as a linear function of the observations $B = \sum m_i Y_i$ and using the assumption of linearity $\mathbb{E}(Y_i) = \alpha + \beta x_i$, show that $\mathbb{E}(B) = \beta$. [*Hint*: $\mathbb{E}(B) = \sum m_i \mathbb{E}(Y_i)$ where $m_i \equiv \dfrac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_i - \bar{x})^2}$]

$E(B) = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$

$= \frac{\Sigma(X_i - \bar{X})Y_i}{\Sigma(X_i - \bar{X})^2}$

$= \frac{1}{\Sigma(X_i - \bar{X})^2} E(\Sigma(X_i - \bar{X})Y_i)$

$= \frac{1}{\Sigma(X_i - \bar{X})^2} \Sigma(X_i - \bar{X})E(Y_i)$

$= \frac{1}{\Sigma(X_i - \bar{X})^2} \Sigma(X_i - \bar{X})E(\alpha + \beta X_i + \varepsilon_i)$

$= \frac{1}{\Sigma(X_i - \bar{X})^2} \Sigma(X_i - \bar{X})(\alpha + \beta X_i + E(\varepsilon_i))$

$= \frac{1}{\Sigma(X_i - \bar{X})^2} \Sigma(X_i - \bar{X})(\alpha + \beta X_i)$

$= \frac{1}{\Sigma(X_i - \bar{X})^2} (\Sigma(X_i - \bar{X})\alpha + \Sigma(X_i - \bar{X})\beta X_i)$

$= \frac{\beta}{\Sigma(X_i - \bar{X})^2} (\Sigma(X_i - \bar{X})X_i)$

$= \beta$

b. Challenge Problem! Demonstrate the unbiasedness of the least-squares estimators A for α in simple regression. Show that A can also be written as a linear function of the Y_is. Then show that $\mathbb{E}(A) = \alpha$.

$E(A) = E(\bar{Y} - B\bar{X})$

$= E(\frac{1}{n}\Sigma Y_i - \frac{\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{\Sigma(X_i-\bar{X})^2}\bar{X})$

$= E(\frac{1}{n}\Sigma Y_i - \frac{\Sigma(X_i-\bar{X})Y_i}{\Sigma(X_i-\bar{X})^2}\bar{X})$

$= E(\Sigma(\frac{1}{n} - \frac{\bar{X}(X_i-\bar{X})}{\Sigma(X_i-\bar{X})^2})Y_i)$

$= E(\Sigma(\frac{1}{n} - \frac{\bar{X}(X_i-\bar{X})}{\Sigma(X_i-\bar{X})^2})(\alpha + \beta X_i + \varepsilon_i))$

$= E(\Sigma(\frac{1}{n} - \frac{\bar{X}(X_i-\bar{X})}{\Sigma(X_i-\bar{X})^2}))E(\alpha + \beta X_i + \varepsilon_i)$

$= E(\Sigma(\frac{1}{n} - \frac{\bar{X}(X_i-\bar{X})}{\Sigma(X_i-\bar{X})^2}))(\alpha + \beta X_i + E(\varepsilon_i))$

$= \Sigma(\frac{1}{n} - \frac{\bar{X}(X_i-\bar{X})}{\Sigma(X_i-\bar{X})^2})(\alpha + \beta X_i)$

$= \Sigma\frac{\alpha}{n} + \Sigma\beta\frac{X_i}{n} - \alpha\frac{\bar{X}\Sigma(X_i-\bar{X})}{\Sigma(X_i-\bar{X})^2} - \beta\frac{\bar{X}\Sigma(X_i-\bar{X})X_i}{\Sigma(X_i-\bar{X})^2}$

$= \alpha + \beta\bar{X} - \alpha * 0 - \beta\bar{X}$

$= \alpha$

c. Challenge Problem! Using the assumptions of linearity, constant variance, and independence along with the fact that A and B can be expressed as a linear function of the Y_is, derive the sampling variances of A and B in a simple regression. [*Hint*: $\mathbb{V}(B) = \sum m_i^2\mathbb{V}(Y_i)$]

$Var(B) = Var(\frac{\Sigma(X_i-\bar{X})Y_i}{\Sigma(X_i-\bar{X})^2})$

$= \frac{1}{(\Sigma(X_i-\bar{X})^2)^2}Var(\Sigma(X_i - \bar{X})Y_i)$

$= \frac{1}{(\Sigma(X_i-\bar{X})^2)^2}Var(\Sigma(X_i - \bar{X})(\alpha + \beta X_i + \varepsilon_i))$

$= \frac{1}{(\Sigma(X_i-\bar{X})^2)^2}Var(\Sigma(X_i - \bar{X}(\alpha + \beta X_i) + \Sigma(X_i - \bar{X})\varepsilon_i)$

$= \frac{1}{(\Sigma(X_i-\bar{X})^2)^2}Var(\Sigma(X_i - \bar{X})\varepsilon_i) = \frac{1}{(\Sigma(X_i-\bar{X})^2)^2}\Sigma Var((X_i - \bar{X})\varepsilon_i)$

$= \frac{1}{(\Sigma(X_i-\bar{X})^2)^2}(\Sigma(X_i - \bar{X})^2 Var(\varepsilon_i))$

$= \frac{1}{(\Sigma(X_i-\bar{X})^2)^2}(\Sigma(X_i - \bar{X})^2\sigma_\varepsilon^2)$

$= \frac{\sigma_\varepsilon^2}{\Sigma(X_i-\bar{X})^2}$

$$
\begin{aligned}
&Var(A) = Var(\bar{Y} - B\bar{X}) \\
&= Var(\bar{Y}) + Var(B\bar{X}) - 2Cov(\bar{Y}, B\bar{X}) \\
&= Var(\bar{Y}) + Var(B\bar{X}) \\
&= Var(\tfrac{\Sigma Y_i}{n}) + \bar{X}^2 Var(B) \\
&= \tfrac{1}{n^2} Var(\Sigma Y_i) + \bar{X}^2 Var(B) \\
&= \tfrac{1}{n^2} \Sigma Var(Y_i) + \bar{X}^2 Var(B) \\
&= \tfrac{1}{n^2} \Sigma Var(\alpha + \beta X_i + \varepsilon_i) + \bar{X}^2 Var(B) \\
&= \tfrac{1}{n^2} \Sigma Var(\varepsilon_i) + \bar{X}^2 Var(B) \\
&= \tfrac{1}{n^2} n\sigma_\varepsilon^2 + \bar{X}(\tfrac{\sigma_\varepsilon^2}{\Sigma(X_i - \bar{X})^2}) \\
&= \sigma_\varepsilon^2(\tfrac{1}{n} + \tfrac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}) \\
&= \sigma_\varepsilon^2(\tfrac{\Sigma(X_i - \bar{X})^2 + n\bar{X}^2}{n\Sigma(X_i - \bar{X})^2}) \\
&= \sigma_\varepsilon^2(\tfrac{\Sigma X_i^2 - 2\bar{X}\Sigma X_i + \Sigma\bar{X}^2 + n\bar{X}^2}{n\Sigma(X_i - \bar{X})^2}) \\
&= \sigma_\varepsilon^2(\tfrac{\Sigma X_i^2 - 2\bar{X}\Sigma X_i + 2n\bar{X}^2}{n\Sigma(X_i - \bar{X})^2}) \\
&= \sigma_\varepsilon^2(\tfrac{\Sigma X_i^2 - 2\frac{\Sigma X_i}{n}\Sigma X_i + 2n\bar{X}^2}{n\Sigma(X_i - \bar{X})^2}) \\
&= \sigma_\varepsilon^2(\tfrac{\Sigma X_i^2 - \frac{2\Sigma X_i^2}{n} + 2n\frac{\Sigma X_i^2}{n^2}}{n\Sigma(X_i - \bar{X})^2}) \\
&= \tfrac{\sigma_\varepsilon^2 \Sigma X_i^2}{n\Sigma(X_i - \bar{X})^2}
\end{aligned}
$$