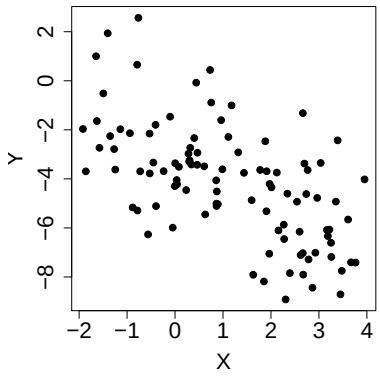
Chicago Booth BUS 41100
Solutions to Midterm Exam **SAMPLE #3**

Instructor: Max H. Farrell

These solutions are a guide only! Your answers should show more work/detail/reasoning.

1 Short Answer & Multiple Choice

(a) Which of the following best describes the least-squares line fit to the data shown in the plot?



(i) $b_0 = 0$, $b_1 = -1$

(ii) $b_0 = -3$, $b_1 = 1$

(iii) $b_0 = -5$, $b_1 = 2$

(iv) $b_0 = -3$, $b_1 = -1$

(v) $b_0 = 0$, $b_1 = -3$

Solution. *Choice (iv) is correct. The slope is clearly negative, ruling out (ii) and (iii). Looking at the axes shows that for a one-unit change in X, Y certainly does not fall by 3, ruling out (v). Choices (i) and (iv) have different intercepts, and following up from $X = 0$, the intercept clearly can not be 0, ruling out (i).*

(b) Suppose you estimate a simple linear regression model and obtain a `t-value` for the slope coefficient of `-3.1`. Based on this, which of the following statements are WRONG? (Circle all that apply.)

(i) A 95% confidence interval for the true slope would exclude 0.

(ii) It is *possible* that the point estimate for the slope is $b_1 = 4$.

(iii) At the 10% level you fail to reject the null hypothesis that the true slope is equal to 0.

(iv) The probability that the true slope is negative is greater than the probability that the true slope is positive.

Solution. *(i) is true* `t-value` *is greater than 2 in absolute value. (ii) is WRONG: the* `t-value` *is* b_1/s_{b_1}*, so it can only be negative if* $b_1 < 0$*. (iii) is WRONG because the 90% CI is narrower (excludes even more values) than the 95% interval. (iv) is WRONG because the true slope is just a number, so it is either positive or negative, there is no probability.*

(c) Suppose we form a 95% confidence interval based on a simple linear regression. Explain in words what is meant by "95%".

Solution. *We mean that if we repeated the same experiment many times, in 95% of them would the confidence interval contain the truth. Or rephrased, we mean that if we draw many different random data sets, and form a confidence interval in each, then 95% of those intervals contain the true parameter. The number 95 itself means that we accept a false positive rate of* $\alpha = 0.05 = 1/20$*.*

(d) Suppose that in the population $Y = 2 - X - 3X^3 + \log(X) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 5X^2)$, so the true conditional expectation is $\mathbb{E}[Y|X] = 2 - X - 3X^3 + \log(X)$. In your sample, you run a simple linear regression of Y on X and obtain $\hat{Y}_i = b_0 + b_1 X_i$ and residuals e_i for each point. Then you will have $\texttt{corr}(X, e) \neq 0$ because the linear regression model is incorrect. TRUE or FALSE? Justify your answer.

Solution. *FALSE. There may be a pattern to the residuals, but running least squares always forces the correlation to be zero.*

(e) We have data on three variables: Y, X_1, and X_2. Then we create a new variable by adding together the two X variables: $X_3 = X_1 + X_2$. We now regress Y on X_1, X_2, and X_3. Explain why we can't find the regression coefficient on X_3 even though $\texttt{corr}(X_1, X_3)$ and $\texttt{corr}(X_2, X_3)$ are not 1.

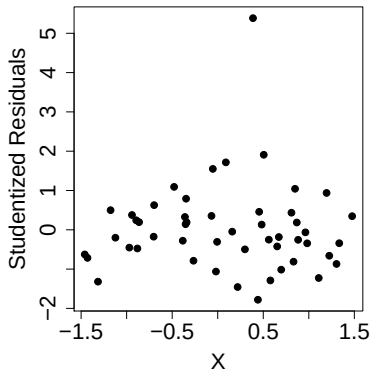
Solution. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3(X_1 + X_2) = \beta_0 + (\beta_1 + \beta_3)X_1 + (\beta_2 + \beta_3)X_2$*, so* X_3 *has no independent linear information.*

(f) We have data on two variables: Y and X. A positive correlation between X and the residuals from a simple linear fit would indicate that the slope is too small relative to the least squares regression line. TRUE or FALSE? Justify your answer.

Solution. *TRUE: a positive correlation means that the residuals too low at first, then too high, so the predictions must be too high and then too low, so the line must be rotated counter-clockwise, i.e. made steeper.*

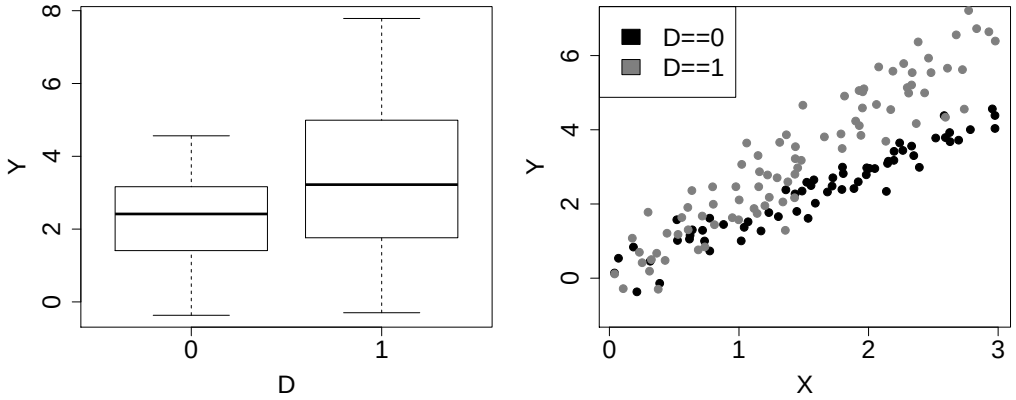
2 Plots & Assumptions

(a) For the plot below, use the space to the right to describe what you think is problematic. Which of our standard linear regression assumptions are violated, if any? If there is a potential problem, what would your next step(s) be?



Solution. *There is a potential outlier: the single point with the very large residuals. The next step would be to investigate this point and see if there is a non-statistical reason to remove it.*

(b) Plotted below are data on 3 variables: a continuous Y and X and a binary variable $D \in \{0, 1\}$.



(i) Based on these two plots, what model specification would you recommend where Y is the outcome and X and D are the independent variables?

Solution. $Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 X \times D + \varepsilon$, *because the second plot clearly shows different slopes. You could argue that removing the* $\beta_1 D$ *term is fine, because it appears that the intercepts are the same, but it is better to leave it in.*

(ii) Based on these two plots, which, if any, of our standard linear regression assumptions are violated?

Solution. *Nonconstant variance in* D.

3 Simple Linear Regression

The Capital Asset Pricing Model relates the returns of an asset, given by R_A, to the returns of the market, R_M through a linear regression equation

$$R_A = \alpha + \beta R_M + \varepsilon.$$

Assume that all our standard simple linear regression assumptions are met. We estimate the regression and obtain

$$R_A = 0.03 - 0.6 R_M + e,$$

where the sample mean of $R_M = 0.5$, the standard error of the intercept is 0.01, the standard error of the slope is 0.25, the sample variance of $R_M = 1$, and the residual standard error is 0.02.

(a) Is the slope coefficient statistically significant at the 0.05 level?

Solution. *Yes, the t-statistic is* $b_1/s_{b_1} = -0.6/0.25$*, which is greater than 2 in absolute value.*

(b) Compute the correlation between R_M and e.

Solution. *Zero, by the definition of least squares regression.*

(c) What is the average return of the asset in this sample?

Solution. *-0.27. Plug in and solve using* $b_0 = \bar{Y} - b_1\bar{X}$*.*

(d) Estimate the variance of R_A.

Solution. $\mathbb{V}[R_A] = \mathbb{V}[0.03 - 0.6R_M + e] = (-0.6)^2\mathbb{V}[R_M] + \mathbb{V}[e] - 2 \times 0.6\text{cov}(R_a, e) = 0.3604$*, because the sample variance of* $R_M = 1$*, the residual standard error is 0.02, and their covariance is zero.*

(e) Estimate the R^2 for this regression.

Solution. $R^2 = 0.999 \approx 1$*.* R^2 *is the squared correlation, which can be found by solving* $b_1^2 = r_{xy}^2 s_y^2/s_x^2$*.*

4 Multiple Linear Regression

Researchers at General Motors collected data on 60 U.S. Standard Metropolitan Statistical Areas (SMSA's) in a study of whether or not air pollution contributes to mortality. The dependent variable for analysis is age adjusted `Mortality` rate. Explanatory variables include two pollution measures and two demographic measures:

`HCPot`	HydroCarbons pollution potential index
`NOxPot`	Nitrous Oxide pollution potential index
`HighSchool`	= 1 if the SMSA median eduction is $\geq$ 12 years, = 0 otherwise
`PopD`	Population density

(a) Consider first a linear regression of `log(Mortality)` on `log(HCPot)`, given by the `summary` below.

```
Call:
lm(formula = log(Mortality) ~ log(HCPot))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.8222     0.0219   311.2   <2e-16 ***
log(HCPot)    0.0079     0.0073     1.1      0.3

Residual standard error: 0.066 on 58 degrees of freedom
Multiple R-squared:  0.02,      Adjusted R-squared:  0.0029
F-statistic: 1.2 on 1 and 58 DF,  p-value: 0.28
```

(i) Provide a complete description of the linear regression model and assumptions that are implicitly used by this output.

Solution. $\texttt{log(Mortality)} = \beta_0 + \beta_1\texttt{log(HCPot)} + \varepsilon$*,* $\varepsilon \sim \mathcal{N}(0, \sigma^2)$*, and everything is independent. Or the equivalent statement in words. See Week 2.*

(ii) Explain in detail what the parameter estimated by the coefficient on `log(HCPot)` represents. Then, in that context, interpret the estimate numerically.

Solution. *See Week 4, this is a log-log model, so* 0.0079 *is the pollution elasticity of mortality. It is not significantly different from zero however.*

(b) We now consider the second pollution measure, `NOxPot`. Consider the following `summary`.

```
Call:
lm(formula = log(Mortality) ~ log(NOxPot))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.8081     0.0178   382.1   <2e-16 ***
log(NOxPot)   0.0156     0.0069     2.3     0.03 *

Residual standard error: 0.064 on 58 degrees of freedom
Multiple R-squared:  0.082,     Adjusted R-squared:  0.066
F-statistic: 5.2 on 1 and 58 DF,  p-value: 0.026
```

Compare to the `summary` from **part (a)**. Which pollution potential is a better predictor of `Mortality` and why? Can you tell by comparing the two regression outputs?

Solution. `NOxPot` *is better: the slope coefficient's p-value is signficant here, so we can not reject the guess that the elasticity is zero. Equivalently (since this is SLR), the F-test is significant here.*

(c) Consider now using both pollution measures in the same regression model.

```
Call:
lm(formula = log(Mortality) ~ log(HCPot) + log(NOxPot))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.850      0.021   328.1   <2e-16 ***
log(HCPot)    -0.064      0.019    -3.3    0.002 **
log(NOxPot)    0.074      0.019     3.9    2e-04 ***

Residual standard error: 0.059 on 57 degrees of freedom
Multiple R-squared:  0.23,      Adjusted R-squared:  0.2
F-statistic: 8.5 on 2 and 57 DF,  p-value: 6e-04
```

(i) Explain in detail what the parameter estimated by the coefficient on `log(HCPot)` represents. Explicitly compare to **part (a)(ii)**. Then interpret the estimate numerically, again comparing to **(a)(ii)**.

Solution. *Still an elasticity, but controlling for* `NOxPot`*, whereas (a)(ii) did not control for* `NOxPot`*. Numerically, we see that for a fixed value of* `NOxPot`*, the* `HCPot` *elasticity is actually negative, so more pollution potential means lower mortality, whereas in (a)(ii) it was positive, though we could not reject the hypothesis of a zero elasticity.*

(ii) Explain the change in the standard errors for the coefficients on `log(HCPot)` (comparing to **part (a)**) and `log(NOxPot)` (comparing to **part (e)**).

Solution. *The standard errors are much larger here, and this is because of multicollinearity. You can't compute it from this output, but in fact, the correlation is about 0.9.*

(d) Let us now consider demographic information, beginning with `HighSchool`, for which a simple linear regression gives the `summary` below.

```
Call:
lm(formula = log(Mortality) ~ HighSchool)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.8602     0.0084   815.5   <2e-16 ***
HighSchoolTRUE  -0.0806     0.0188    -4.3    7e-05 ***

Residual standard error: 0.058 on 58 degrees of freedom
Multiple R-squared:  0.24,      Adjusted R-squared:  0.23
F-statistic:  18 on 1 and 58 DF,  p-value: 6.9e-05
```

(i) Provide a complete description of the linear regression model and assumptions that are implicitly used by this output.

Solution. *See Week 3, just differences in means.*

(ii) Explain in detail what the parameter estimated by the coefficient on `HighSchool` represents. Then, in that context, interpret the estimate numerically.

Solution. *See Week 3. Numerically, SMSA's with a median education level at least HS have 8% lower mortality.*

(e) Consider the `summary` below.

```
Call:
lm(formula = log(Mortality) ~ HighSchool * log(HCPot))

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                  6.757      0.020   339.7   <2e-16 ***
HighSchoolTRUE               0.030      0.043     0.7      0.5
log(HCPot)                   0.041      0.007     5.5    9e-07 ***
HighSchoolTRUE:log(HCPot)   -0.043      0.012    -3.6    7e-04 ***

Residual standard error: 0.05 on 56 degrees of freedom
Multiple R-squared:  0.5,       Adjusted R-squared:  0.5
F-statistic: 2e+01 on 3 and 56 DF,  p-value: 1e-08
```

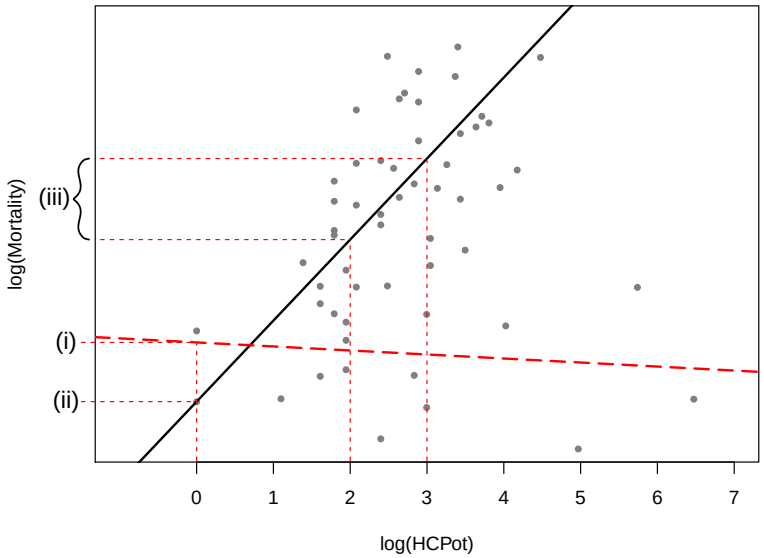
(i) Provide a complete description of the linear regression model and assumptions that are implicitly used by this output.

Solution. *See Week 3, different slopes and different intercepts.*

(ii) Explain in detail what the parameter estimated by the coefficient on `HighSchoolTRUE:log(HCPot)` represents. Then, in that context, interpret the estimate numerically.

Solution. *See Week 3. This is the additional slope for the SMSA's that have higher education. Numerically, the slope is actually lower, because* $0.041 + -0.043 = -0.002$*, so in SMSA's with median education at least high school,* `HCPot` *has a slightly negative effect, but basically zero.*

(f) Consider the plot below which is based on the regression reported in **part (e)**.



Explain briefly what is shown in this plot. Describe both what information the plot represents and what you conclude from it.

Solution. *See Week 3, different slopes and different intercepts We conclude that* `HCPot` *has a very different relationship to mortality between the two types. However, we are also worried about the very small number of points in the bottom right of the graph, which have high pollution, but also low mortality, and, because the* `HighSchoolTRUE` *line goes through them, these are more educated SMSA's. This is likely driving the result that the slope is negative for this group.*

Using the `summary` output from **part (e)** above, compute the numeric values for the three points labeled on the graph as:

Solution. *(i) = 6.757 + 0.030; (ii) = 6.757; (iii) = 0.041.*

(g) Now consider adding `PopDensity` to the regression in **part (e)**, which yields this `summary`.

```
Call:
lm(formula = log(Mortality) ~ HighSchool * log(HCPot) + log(PopDensity))

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                  6.664      0.153    43.5   <2e-16 ***
HighSchoolTRUE               0.025      0.044     0.6    0.574
log(HCPot)                   0.039      0.009     4.5    4e-05 ***
log(PopDensity)              0.012      0.020     0.6    0.542
HighSchoolTRUE:log(HCPot)   -0.041      0.013    -3.2    0.002 **

Residual standard error: 0.05 on 55 degrees of freedom
Multiple R-squared:  0.5,       Adjusted R-squared:  0.5
F-statistic: 1e+01 on 4 and 55 DF,  p-value: 4e-08
```

Based on this output, should we add `log(PopDensity)` to the regression? Why or why not? Cite specific numerical values from the output to support your argument.

Solution. *The partial F test says no: the partial F test p-value would be 0.542.*

5 Logistic Regression

For 546 homes we observe the following informaiton

`AC` = 1 if it has air conditioning, 0 if not,
`price` = sale price, and
`bedrooms` = number of bedrooms.

Our goal is to predict which houses have air condition and which do not.

(a) First we use a logistic regression with only `bedrooms` and obtain the following output.

```
Call:
glm(formula = AC ~ bedrooms, family = "binomial")

Coefficients:
            Estimate Std. Error z value    Pr(>|z|)
(Intercept)   -2.179      0.399   -5.46 0.000000047
bedrooms       0.469      0.127    3.68     0.00023
```

Give a precise, numerical interpretation of the relationship between `bedrooms` and `AC` by:

(i) Interpreting the coefficient estimate, 0.469, in this context.
Solution. *See Week 4.*

(ii) Interpreting the statistical significance as best you can with this output.
Solution. *The most useful thing to do is probably the usual thing: test the null of $\beta_1 = 0$, which we reject because $0.469/0.127 > 2$. So we conclude that at the 5% level bedrooms are associated with an increase in the probability of having AC.*

(b) We now switch to using `price` and obtain the following output.

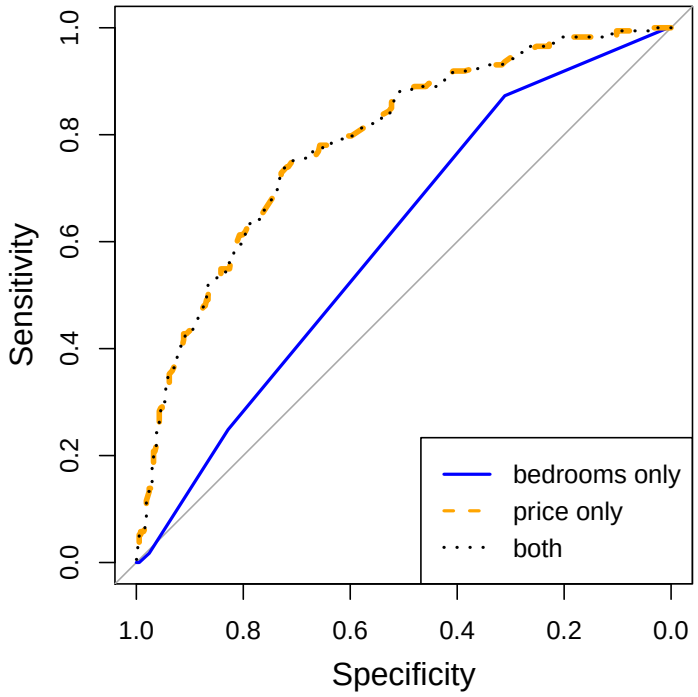
```
Call:
glm(formula = AC ~ price, family = "binomial")

Coefficients:
              Estimate  Std. Error z value            Pr(>|z|)
(Intercept) -3.76328283  0.34599029   -10.9 <0.0000000000000002
price        0.00004223  0.00000459     9.2 <0.0000000000000002
```

Based on this output, is `price` or `bedrooms` better at predicting whether or not a house has AC? Cite specific numeric evidence in favor or your argument.

Solution. *There is really no way to tell from this output. Both are statistically significant.*

Use the plot below to answer the next two questions.



(c) Provide an intuitive explanation of what is being plotted here.

Solution. *See Week 4.*

(d) Based on this plot, comment on how `bedrooms` contribute to predicting `AC` when used in conjunction with `price`.

Solution. *We can see that* `bedrooms` *adds nothing to the model's ability to classify houses into AC yes/no.*

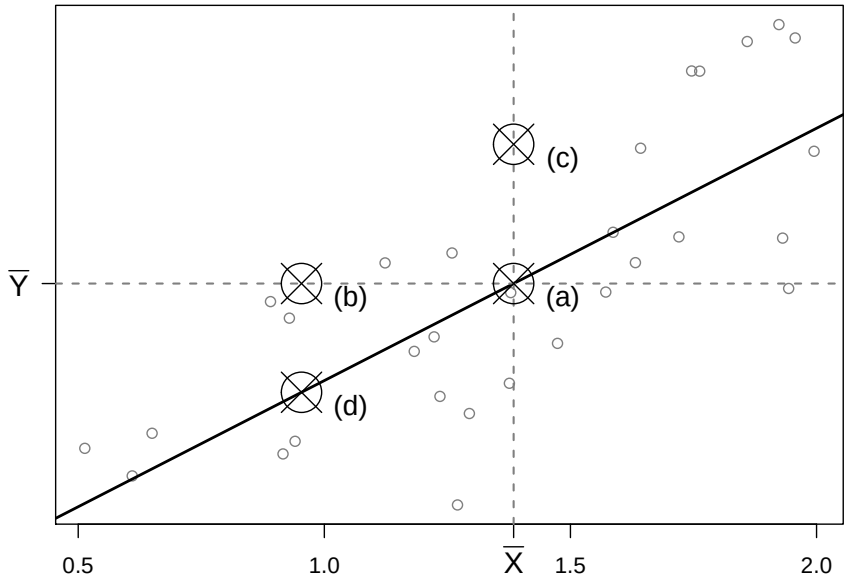
6 Understanding Simple Linear Regression

This question is about how parameter estimates change as the data changes, but the underlying model is fixed. Throughout, assume that $Y = \beta_0 + \beta_1 X + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The parameters β_0, β_1, and σ^2 are fixed but our estimates of them change as the data changes. We start with an i.i.d. sample of $n = 30$ points and obtain the following `summary`:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.81       0.57     3.2    0.004 **
X               2.59       0.39     6.6    4e-07 ***

Residual standard error: 0.91 on 28 degrees of freedom
Multiple R-squared:  0.61,      Adjusted R-squared:  0.6
F-statistic:  44 on 1 and 28 DF,  p-value: 3.6e-07
```

The 30 data points (grey circles) and the least squares fit (black line) are plotted below. **One at a time** we add a new data point at one of the four spots labeled **(a)**, **(b)**, **(c)**, and **(d)**. X marks the spot! E.g., point **(a)** is at $(\bar{X}, \bar{Y})$.



Using the plot and the `summary` above, indicate the effect of adding the new data point on each parameter estimate by circling **exactly one** of {up / dn / same / ?} in each cell of the table below. Circle "up" if the estimate will be higher with the new point, "dn" if it will be lower, "same" for unchanged, and "?" if the effect is uncertain given the information.

Parameter	New Data Point			
	(a)	**(b)**	**(c)**	**(d)**
b_0	up / dn / (same) / ?	(up) / dn / same / ?	(up) / dn / same / ?	up / dn / (same) / ?
$\texttt{se}(b_0)$	up / (dn) / same / ?	up / dn / same / (?)	up / dn / same / (?)	up / (dn) / same / ?
b_1	up / dn / (same) / ?	up / (dn) / same / ?	up / dn / (same) / ?	up / dn / (same) / ?
$\texttt{se}(b_1)$	up / (dn) / same / ?	up / dn / same / (?)	up / dn / same / (?)	up / (dn) / same / ?

Justify/explain your answers for point **(c)**:

Solution. *Let's go through all the basic ingredients before putting them together.*

- *n has increased by 1.*
- *$\bar{X}$ does not change since the new point is exactly at the old average.*
- *$\bar{Y}$ increases.*
- *Sample variance of X: Instead of $s_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ let's just use the summation part: $\sum_{i=1}^n (X_i - \bar{X})^2$ has not changed because the new point is exactly at the old average.*
- *Sample covariance: Instead of s_{XY} consider just the summation part: $\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$, which also has not changed because the new point is exactly at the old average for X.*
- *The only piece left is the residual standard error s, or s^2. It is not known how this changes: if point (c) is very "far" from the line, then s^2 will increase because the squared new residual $e_{(c)}^2$ will be a bigger increase than the offsetting increase in the sample size, going from $30 - 2 \to 31 - 2$, but if point (c) is close to the line then the logic flips and s^2 would fall. There's no way to known from the information given.*

Putting these together:

- *$b_0 = \bar{Y} - b_1\bar{X}$, so this goes up since b_1 does not change.*
- *$\texttt{se}(b_0)$ is unknown because s^2 is unknown, even though the increase in n would otherwise lead to a decrease and the fact that $\bar{X}^2$ and $\sum_{i=1}^n (X_i - \bar{X})^2$ are the same would indicate no change.*
- *$b_1 = \dfrac{s_{XY}}{s_X^2} = \dfrac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$. Neither of these change, neither does b_1.*
- *$\texttt{se}(b_1)$: same logic as $\texttt{se}(b_0)$.*

Bonus solution for point (b). *The standard errors are uncertain because the change in s^2, relative to anything else, is unknown from this information. The reason that b_0 increases and b_1 falls is that the sample mean point $(\bar{X}, \bar{Y})$ has to lie on the line and so in order for the regression line to move closer to (b), which it must when it tries to minimize the new sum of squared errors, the intercept must rise, which then implies that the slope falls in order to still go through $(\bar{X}, \bar{Y})$. Imagine other cases. Suppose the intercept stayed the same: then the slope would stay the same too, in order to still go through $(\bar{X}, \bar{Y})$, but that would not minimize the sum of squared errors, since*

now there is a new residual that was not taken into account. If the intercept actually went down, then the slope would have to increase, and this new line would be even further from the point (b), which of course can't happen when we minimize the new sum of squared errors.

Consider your answers for points **(a)** and **(d)**. For each parameter estimate:

- If you marked **different** answers, explain why.

  [*e.g. the slope estimate goes up in (a) but down in (d) because ....*]

- If you marked **identical** answers, explain why. In addition, if you marked "up" or "dn", explain which change would be larger, or that the changes would be the same, and why.

  [*e.g. $se(b_0)$ stays the same in both because ....*]

  [*e.g. $se(b_0)$ goes up in both, but increases by more in (a) because ....*]

  Solution. *The simplest explanation for why the slope and intercept estimate don't change is that for either new point, the residual for the new point relative to the original regression line is zero, i.e. the are both right on the line, so the sum of squared errors is not changed, and there is no need to move the line to get a smaller sum of squared errors. For point (a) the formulas for b_0 and b_1 give the answer immediately, but it is much harder to use them for point (d).*
  As to the standard errors, they go down in (a) because the sample size goes up and all the other ingredients stay the same; the explanation is the same as for point (c) above. But they go down **more** *for point (d) because $\sum_{i=1}^{n}(X_i - \bar{X})^2$ increases in (d) but stays the same in (a) and $\bar{X}^2$ decreases in (d) but stays the same in (a).*