DSAN 5300 Quiz-3 Study Guide

For this study guide, we've assembled a set of practice problems representing three topics covered in Week 2, which will correspond to the three problems on Quiz-3 itself! So, a straightforward way to study would be to:

1. Choose a topic
2. Choose one example problem from that topic and attempt to solve it yourself
3. If you run into trouble solving it on your own – e.g., you find yourself having to look at the provided solutions multiple times – you can then focus on what went wrong (or ask us/your fellow students in the Google Space!), and re-try with another example problem[1]

Topic-1: Estimating coefficients via OLS

Problem-1.1: Intercept-only model

Consider estimating a (very restricted) linear regression model where the sole parameter β_0 represents an "intercept":

$$Y = \beta_0$$

As this written-out form makes clear, this model is basically asking *"If I am restricted only to **horizontal lines** (in the Cartesian xy-plane), which of these horizontal lines 'best fits' the data?"* In other words, if I have to make the same prediction $\hat{y}_i$ for every observation i, what number would provide the least-bad estimate given the dataset $\mathfrak{D} = ((x_1, y_1), \ldots, (x_n, y_n))$?

On the basis of this model, derive the **ordinary least-squares (OLS) estimate** for the sole parameter β_0, as a function of the data inputs $((x_1, y_1), \ldots, (x_n, y_n))$.

For this and all example problems in this section, it's important to notice when the terms you derive can be interpreted in an intuitive way, like "ah, this is the mean of the y_i values!" To this end, your estimate should use the following notation, which denotes the mean x value as $\overline{x}$ and the mean y value as $\overline{y}$.[2]

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

[1] I try my best to avoid skipping "small" steps in the mathematical derivations in Topic-1. So, even though the math in this first part may look daunting due to the long chains of equations, some steps are just simple operations like "divide both sides by 2", which some textbooks skip over but I include just in case!

[2] Concretely: if your solution has the term $\frac{1}{n}\sum_{i=1}^{n} x_i$ somewhere, for example, you should replace this term with just $\overline{x}$!

Solution:

The OLS approach here means finding the value β_0^* for the parameter β_0 which **minimizes the mean squared difference** between the predictions $\widehat{y}(x_i)$ made by our model and the true values y_i:

$$\beta_0^* = \underset{\beta_0}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (\widehat{y}(x_i) - y_i)^2 \right].$$

Since our model estimates predictions for y from values of x via $\widehat{y}(x_i) = \beta_0$, we plug this RHS expression (just β_0 in this case) in for $\widehat{y}$ above to obtain:

$$\beta_0^* = \underset{\beta_0}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (\beta_0 - y_i)^2 \right]$$

And now we can take a derivative of this objective function and set it equal to zero to find this optimal value β_0^*! Starting off:

$$\frac{\partial}{\partial \beta_0} \left[\sum_{i=1}^{n} (\beta_0 - y_i)^2 \right] = 0 \iff \sum_{i=1}^{n} \left[\frac{\partial}{\partial \beta_0} (\beta_0 - y_i)^2 \right] = 0$$

Linearity of Derivative Operator

In the previous step, we were able to "move" the derivative operator inside the sum because of the **linearity of the derivative operator**: that is, because this operator satisfies the general property:

$$\frac{\partial}{\partial x}[f(x) + g(x)] = \frac{\partial}{\partial x} f(x) + \frac{\partial}{\partial x} g(x)$$

Continuing with our derivation,

$$\begin{aligned} &\sum_{i=1}^{n} \left[\frac{\partial}{\partial \beta_0} (\beta_0 - y_i)^2 \right] = 0 \iff \sum_{i=1}^{n} 2(\beta_0 - y_i) = 0 \\ \iff &\sum_{i=1}^{n} (\beta_0 - y_i) = 0 \iff \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} y_i = 0 \\ \iff &n\beta_0 = \sum_{i=1}^{n} y_i \iff \beta_0 = \frac{1}{n} \sum_{i=1}^{n} y_i \\ \iff &\boxed{\beta_0^* = \overline{y}} \end{aligned}$$

And thus (as you may have guessed intuitively, but you've now derived mathematically), if your model is restricted to the point that you have to guess a **single fixed number** for every prediction $\widehat{y}_i$, the best number to choose is the **mean of the y_i values!**

Problem-1.2: Slope-only model

In this problem we again consider estimating a "restricted" linear regression model, where in this case the sole parameter β_1 represents the **slope** of a line passing through the origin[3]:

$$Y = \beta_1 X$$

As this written-out form makes clear, this model is basically asking *"If I am restricted only to* ***lines passing through the origins*** *(in the Cartesian xy-plane), which of these lines 'best fits' the data?"*

Recall from class how, the "true" data-generating process may not even be in the space of possible lines through the origin! As is the case in Figure 1.

Nonetheless, just as Ptolemy's model could produce good predictions despite being wrong in a deeper sense, here we hope to find the line-through-the-origin that produces the best possible predictions, **regardless** of whether the true DGP is a line passing through the origin! (Concretely, although none of the grey lines in)

On the basis of this model, derive the **ordinary least-squares (OLS) estimate** for the sole parameter β_1, as a function of the data inputs $((x_1, y_1), \ldots, (x_n, y_n))$.

Solution:

The OLS approach here means finding the value β_1^* for the parameter β_1 which **minimizes the mean squared difference** between the predictions $\widehat{y}(x_i)$ made by our model and the true values y_i:

$$\beta_0^* = \underset{\beta_0}{\text{argmin}} \left[\sum_{i=1}^{n} (\widehat{y}(x_i) - y_i)^2 \right].$$

Since our model estimates predictions for y from values of x via $\widehat{y}(x_i) = \beta_1 x_i$, we plug this RHS expression $(\beta_1 x_i)$ in for $\widehat{y}_i$ above to obtain:

$$\beta_1^* = \underset{\beta_1}{\text{argmin}} \left[\sum_{i=1}^{n} (\beta_1 x_i - y_i)^2 \right]$$

[3] If you remember slope-intercept form, you can think of this model as $y = \beta_1 x + b$, but where b is set to 0, hence the zero-intercept property, which implies that the line will pass through the origin $(0, 0)$.

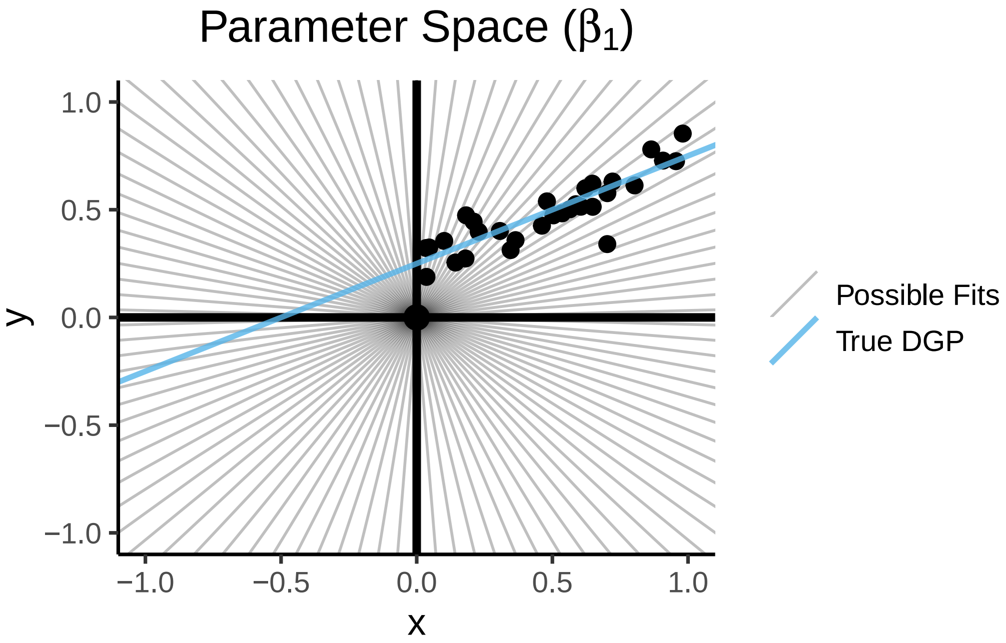

Figure 1: A space of possible fitted models (the space of all lines passing through the origin — an illustrative subset of these possible lines is drawn in gray), where although none of the possible lines perfectly match the true DGP (the blue line), we can still find one line from among the range of gray lines that *best fits* the points in the upper-right quadrant.

And now we can take a derivative of this objective function and set it equal to zero to find this optimal value β_1^*!

$$
\begin{aligned}
&\frac{\partial}{\partial \beta_1}\left[\sum_{i=1}^{n}(\beta_1 x_i - y_i)^2\right] = 0 \\
\iff &\sum_{i=1}^{n}\left[\frac{\partial}{\partial \beta_1}(\beta_1 x_i - y_i)^2\right] = 0 \\
\iff &\sum_{i=1}^{n} 2(\beta_1 x_i - y_i)(x_i) = 0
\end{aligned}
$$

Chain Rule

Here it's important to note that we've been using the **chain rule**,

$$\frac{\partial}{\partial x} g(h(x)) = g'(h(x))h'(x)$$

to handle the fact that the "inner" true-minus-predicted difference $\widehat{y}_i - y_i$ is "nested" within the squaring function $g(x) = x^2$.

In Problem-1.1 we had the "easy case", where the derivative of the inner difference with respect to the parameter β_0 was just the constant 1. Here the derivative of the inner term $(\beta_1 x_i - y_i)$ with respect to the parameter β_1 is **not** the ignorable multiplicative constant 1, but instead $\frac{\partial}{\partial \beta_1}(\beta_1 x_i - y_i) = x_i$, hence the parenthesized x_i term that appears on the LHS in the previous step!

Continuing onto the algebra,

$$
\begin{aligned}
&\sum_{i=1}^{n} 2(\beta_1 x_i - y_i)(x_i) = 0 \iff 2\sum_{i=1}^{n}(\beta_1 x_i - y_i)(x_i) = 0 \\
\iff &\sum_{i=1}^{n}(\beta_1 x_i - y_i)(x_i) = 0 \iff \sum_{i=1}^{n}(\beta_1 x_i^2 - y_i x_i) = 0 \\
\iff &\beta_1 \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i = 0 \iff \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \\
\iff &\boxed{\beta_1^* = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i x_i}}
\end{aligned}
$$

Re-writing the denominator as $\sum_{i=1}^{n} x_i x_i$ instead of $\sum_{i=1}^{n} x_i^2$, as we did above, helps for building some intitial rough intuition around this solution. Think about what happens to this expression for β_1^* in the following cases:

First, when x_i is always **equal to** y_i for a given observation i — i.e., when X and Y are **perfectly positively correlated** (learning the value x_i means you also immediately know the value y_i, and vice-versa) — this expression becomes

$$\beta_1^* = \frac{\sum_{i=1}^n x_i x_i}{\sum_{i=1}^n x_i x_i} = 1$$

On the other hand, if y_i is always the **negative** of x_i (when $x_i = -y_i$ for all observations i) — i.e., when X and Y are **perfectly negatively correlated** (learning the value x_i means you also immediately know the value $y_i = -x_i$, and vice-versa) — the expression becomes

$$\beta_1^* = \frac{\sum_{i=1}^n x_i(-x_i)}{\sum_{i=1}^n x_i x_i} = \frac{(-1)\sum_{i=1}^n x_i x_i}{\sum_{i=1}^n x_i x_i} = -1$$

Lastly, if y_i is **always 0** no matter what the corresponding x_i value is, so that there is **no correlation** whatsoever between X and Y (learning the value x_i gives you no information about the value y_i, and vice-versa), the expression becomes

$$\beta_1^* = \frac{\sum_{i=1}^n x_i(0)}{\sum_{i=1}^n x_i x_i} = \frac{0}{\sum_{i=1}^n x_i x_i} = 0$$

...Food for thought!

Problem-1.3: Fixed-slope model

This model builds on the model from the previous problem, in that we still only have **one** parameter to estimate, β_0. In this case, however, we introduce a slope on X as well, but a slope which is **fixed** to some **specific number (constant)** c:

$$Y = \beta_0 + cX$$

To minimize confusion, you can think of the estimation of this model in two “stages”: Someone comes along and **gives you** a value for c first (maybe $c = 5$, or $c = \pi$, etc... some fixed numeric value), and **then** your job now is to use the OLS approach to estimate the optimal parameter value β_0^* given this fixed c value.

Solution:

We want to find the value β_0^* for the parameter β_0 that minimizes the mean squared difference between the predictions $\widehat{y}(x_i)$ and the true values y_i:

$$\beta_0^* = \min_{\beta_0} \left[\sum_{i=1}^{n} (\widehat{y}(x_i) - y_i)^2 \right].$$

Since our model estimates predictions for y from values of x via $\widehat{y}(x_i) = \beta_0 + cx_i$, we plug this in for $\widehat{y}_i$ above to obtain:

$$\beta_0^* = \min_{\beta_0} \left[\sum_{i=1}^{n} ((\beta_0 + cx_i) - y_i)^2 \right]$$

And now we can take a derivative of this objective function and set it equal to zero to find this optimal value β_0^*!

$$\begin{aligned}
&\frac{\partial}{\partial \beta_0} \left[\sum_{i=1}^{n} (\beta_0 + cx_i - y_i)^2 \right] = 0 \iff \sum_{i=1}^{n} \left[\frac{\partial}{\partial \beta_0} (\beta_0 + cx_i - y_i)^2 \right] = 0 \\
&\iff \sum_{i=1}^{n} 2(\beta_0 + cx_i - y_i) = 0 \iff \sum_{i=1}^{n} \beta_0 + \sum_{i=1}^{n} cx_i - \sum_{i=1}^{n} y_i = 0 \\
&\iff n\beta_0 = \sum_{i=1}^{n} y_i - c \sum_{i=1}^{n} x_i \iff \beta_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - c \cdot \frac{1}{n} \sum_{i=1}^{n} x_i \\
&\iff \boxed{\beta_0^* = \overline{y} - c \cdot \overline{x}}
\end{aligned}$$

We could take time to try and build intuition about this particular solution, but we'll see that it really emerges as just a "special case" of the more general intuition in Problem-1.5 below.

Problem-1.4: Fixed-intercept model

Consider a model with a similar overall setup to Problem-1.2, but where we now have a **fixed, constant intercept** c, and a single **parameter** we'd like to estimate, β_1, the **slope** on X:

$$Y = c + \beta_1 X$$

Use the OLS approach to find the optimal parameter value β_1^* in this case.

Solution:

As with the above two problems, our goal in using the OLS approach is to derive an estimate β_1^* for the model parameter β_1 which **minimizes the squared difference** between true values y_i and model predictions $\widehat{y}_i(x_i)$:

$$\beta_1^* = \underset{\beta_1}{\text{argmin}} \left[\sum_{i=1}^{n} (\widehat{y}(x_i) - y_i)^2 \right]$$

Since the model in this problem produces estimates $\widehat{y}_i$ for a given observation i as $\widehat{y}_i = c + \beta_1 x_i$, we plug this RHS in for the $\widehat{y}(x_i)$ term above to obtain:

$$\beta_1^* = \underset{\beta_1}{\text{argmin}} \left[\sum_{i=1}^{n} (c + \beta_1 x_i - y_i)^2 \right]$$

And now we take the derivative of the minimand (the term inside the square brackets), set this derivative equal to zero, and solve the resulting equation for β_1 to obtain the OLS solution β_1^*!

$$\begin{aligned}
&\frac{\partial}{\partial \beta_1} \left[\sum_{i=1}^{n} (c + \beta_1 x_i - y_i)^2 \right] = 0 \iff \sum_{i=1}^{n} \frac{\partial}{\partial \beta_1} (c + \beta_1 x_i - y_i)^2 = 0 \\
\iff &\sum_{i=1}^{n} 2(c + \beta_1 x_i - y_i)(x_i) = 0 \iff 2\sum_{i=1}^{n} (c x_i + \beta_1 x_i^2 - x_i y_i) = 0 \\
\iff &\sum_{i=1}^{n} (c x_i + \beta_1 x_i^2 - x_i y_i) = 0 \iff \sum_{i=1}^{n} c x_i + \beta_1 \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i = 0 \\
\iff &\beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} c x_i \iff \beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} c x_i}{\sum_{i=1}^{n} x_i^2} \\
\iff &\boxed{\beta_1^* = \frac{\sum_{i=1}^{n} x_i (y_i - c)}{\sum_{i=1}^{n} x_i^2}}
\end{aligned}$$

As in the previous problem, we could derive intuition for this solution, but in fact the easiest way to interpret this is: it's the solution we would obtain if we just "shifted" all y_i values by an amount c and then used the **slope only** model from Probem-1.2 above!

In other words, the above boxed solution is identical to the Problem-1.2 solution, if we take that solution and replace y_i everywhere by $y_i - c$. This is easy to see in hindsight, since we could just rewrite our model in this Problem by subtracting c from both sides to obtain

$$Y - c = \beta_1 X,$$

and then if we define $\tilde{Y} = Y - c$, we fully "re-obtain" the model from Problem-1.2:

$$\tilde{Y} = \beta_1 X$$

Problem-1.5: Full Simple Linear Regression (SLR) model

We've finally made it to the full simple linear regression model introduced in class:

$$Y = \beta_0 + \beta_1 X$$

Use the OLS approach to derive estimates for **both** parameters of this model: the intercept β_0 and the slope β_1.

Solution:

Here we'll *start* working through the admittedly messy/daunting derivation using the same approach we used in each previous Problem, which is important to understand since it will be our first example of optimizing over **multiple** parameters (β_0 and β_1) simultaneously. But, once we reach a certain point where we know that we *could* derive the solution by doing a ton of algebra, we'll switch to a different approach showing how you can arrive at this solution in a much simpler way by using more succinct **matrix algebra** notation.

Full Derivation:

As in all of the earlier cases, we want to find a closed-form expression for the optimal values of the model parameters β_0 and β_1 as a function of the input data $\mathfrak{D} = ((x_1, y_1), \ldots, (x_n, y_n))$.

Since the OLS approach implies (by its name) that we should choose these optimal β_0 and β_1 values so as to **minimize the squares of the prediction errors**, we have

$$\boldsymbol{\theta}^* = (\beta_0^*, \beta_1^*) = \operatorname*{argmin}_{\beta_0, \beta_1} \left[\sum_{i=1}^{n} (\widehat{y}(x_i) - y_i)^2 \right] = \operatorname*{argmin}_{\beta_0, \beta_1} \left[\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2 \right]$$

where the new symbol $\boldsymbol{\theta}$ (the Greek letter "theta") is common notation for the **vector** of **model parameters**: in this case, we take the two parameters β_0 and β_1 and bundle them together to form $\boldsymbol{\theta}$.

First-Order Conditions

In this problem we have to be a bit more specific than *"take the derivative and set it equal to zero"*, since there are now **two** separate derivatives that we care about: the derivative of the loss with respect to β_0 and with respect to β_1.

To handle this complication, we generalize from "setting the derivative equal to zero" to "establishing **first order conditions**": statements that must be true if given values of β_0 and β_1 are indeed "critical points" (possible minimum points). This lets us easily "filter out" non-optimal values, since a given pair of values that does **not** satisfy first-order conditions **cannot** represent the optimal loss-minimizing parameter values[4].

In this case, our two first-order conditions "encode" the criteria that, at a minimum point of our loss function, its derivative with respect to **both** β_0 **and** β_0 will be equal to zero:

$$(\beta_0^*, \beta_1^*) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} L(\beta_0, \beta_1; \mathfrak{D}) \implies \boxed{\frac{\partial L}{\partial \beta_0} = 0} \text{ and } \boxed{\frac{\partial L}{\partial \beta_1} = 0}$$

Using the vector notation from last week's quiz, which we'll pick back up below, this can all be summarized more simply as saying that the first-order condition is satisfied if the **gradient** $\nabla L(\beta_0, \beta_1; \mathfrak{D})$ with respect to the parameter vector (β_0, β_1) is equal to the **zero vector** $(0, 0)$. This phrasing makes it more clear that this is the multivariable generalization of "set the derivative equal to zero": we just replace the term "derivative" with "gradient".

For now, we compute **both** of the partial derivatives that form the first-order condition, then set both equal to zero and solve as a system of **two equations** in **two unknowns**:

$$\begin{aligned}
&\frac{\partial L}{\partial \beta_0} = 0 \iff \frac{\partial}{\partial \beta_0}\left[\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - y_i)^2\right] = 0 \\
&\iff \sum_{i=1}^{n} \frac{\partial}{\partial \beta_0}(\beta_0 + \beta_1 x_i - y_i)^2 = 0 \iff \sum_{i=1}^{n} 2(\beta_0 + \beta_1 x_i - y_i) = 0 \\
&\iff \sum_{i=1}^{n} \beta_0 + \sum_{i=1}^{n} \beta_1 x_i - \sum_{i=1}^{n} y_i = 0 \iff n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0 \\
&\iff n\beta_0 = \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i \iff \beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i \\
&\iff \boxed{\beta_0^* = \overline{y} - \beta_1^* \overline{x}}
\end{aligned}$$

Now that we have a formula for **deriving** the optimal value β_0^* **from** the other optimal value β_1^*, we'll start deriving that by working through our other first-order condition:

$$\begin{aligned}
&\frac{\partial L}{\partial \beta_1} = 0 \iff \frac{\partial}{\partial \beta_1}\left[\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - y_i)^2\right] = 0 \\
&\iff \sum_{i=1}^{n} \frac{\partial}{\partial \beta_1}(\beta_0 + \beta_1 x_i - y_i)^2 = 0 \iff \sum_{i=1}^{n} 2(\beta_0 + \beta_1 x_i - y_i)(x_i) = 0
\end{aligned}$$

[4] The "could be" is awkward but necessary because, as you may have learned in earlier calculus classes, in general multivariable settings the fact that (β_0, β_1) satisfies the first-order conditions is **necessary** but **not sufficient** for optimality. In general you also need to check **second-order conditions**, to confirm whether certain "critical values" are actually minima/maxima and not e.g. "saddle points", and **boundary conditions**, as the minima/maxima may lie at the "ends" of a loss function's domain where its derivative is not necessarily zero. Here, since our quadratic OLS loss function is **convex** (lines between any two points on the curve lie above it) and **unbounded** (ranges from $-\infty$ to ∞), the first-order conditions are sufficient.

$$\iff \sum_{i=1}^{n}(\beta_0 x_i + \beta_1 x_i^2 - y_i x_i) = 0 \iff \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i x_i = 0.$$

At this point, we could try to simplify more, or we could use the expression $\beta_0 = \overline{y} - \beta_1 \overline{x}$ we derived above, plug in $\overline{y} - \beta_1 \overline{x}$ for β_0 in our first-order condition, and derive a closed-form solution (meaning, a solution which does not depend on β_0) for β_1. With that closed-form solution for β_1 in hand, we could complete the problem by plugging this closed-form solution for β_1 **back into** our $\beta_0 = \overline{y} - \beta_1 \overline{x}$ equation and solving to obtain a closed-form solution for β_0 as well.

So, given that we know you're capable of these last algebraic steps (and, they take too long for us to put this question on a Quiz anyways!), a better use of time would be to instead consider the **matrix algebra** representation mentioned above! It will be important to understand this alternative (and more powerful) notation as the class progresses, so take a look at the online Appendix to this study guide, if you have time after going through Topic-2 and Topic-3!

Topic-2: Interpreting SLR Output

The following table shows the output from a regression of `sales` on `newspaper` (both variables from ISLP's `Advertising.csv` dataset, which is introduced in Chapter 1 and discussed throughout Chapter 2), estimated using Python's `statsmodels` library.

```python
import pandas as pd
import statsmodels.formula.api as smf
ad_df = pd.read_csv("Advertising.csv")
lr_model = smf.ols('sales ~ newspaper', data=ad_df)
lr_model_summary = lr_model.fit().summary(slim=True)
lr_model_summary.extra_txt = ''
print(lr_model_summary)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.052
Model:                            OLS   Adj. R-squared:                  0.047
No. Observations:                 200   F-statistic:                     10.89
Covariance Type:            nonrobust   Prob (F-statistic):            0.00115
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     12.3514      0.621     19.876      0.000      11.126      13.577
newspaper      0.0547      0.017      3.300      0.001       0.022       0.087
==============================================================================
```

The unit for all variables in this part (`sales`, `newspaper`, `TV`, and `radio`) is $1,000$ USD, so that for example a value of `51.1234` for `sales` would represent $51,123.40$ USD in sales for a given period of time. Use this output to answer the following questions:

Question-2.1

Which of the following represents the model — the hypothesized relationship between `sales` and `newspaper` — that the above output provides estimates for?

(a) $\texttt{newspaper} = \beta_0 + \beta_1 \cdot \texttt{sales}$
(b) $\texttt{sales} = \beta_0 + \texttt{newspaper}$
(c) $\texttt{sales} = \beta_1 \cdot \texttt{newspaper}$
(d) $\texttt{sales} = \beta_0 + \beta_1 \cdot \texttt{newspaper}$

Solution: (d) $\texttt{sales} = \beta_0 + \beta_1 \cdot \texttt{newspaper}$

Option (a) is a model representing the "opposite" of the model in this Topic: it would provide estimates for the changes in **`newspaper`** associated with changes in **`sales`**. This is not what we want (given the discussion in the book of the purpose of the dataset), since we're hoping to advise a company on how its **sales** increase or decrease for different levels of investment in newspaper ads.

Option (b) has no parameter for the **slope** on `newspaper`, meaning that this model would be pre-assuming a slope of 1 implicitly, meaning that the model would not be flexible enough to model cases where an increase of $1,000$ USD in newspaper ads is associated with an increase in sales of some amount besides $1,000$ USD.

Option (c) has no parameter for the **intercept** on `newspaper`, meaning that this model would be pre-assuming that sales are 0 USD whenever newspaper ad investments are 0 USD. This is probably not an assumption we want to make, since there are many other "mechanisms" besides newspaper ads that may lead to sales (for example, word-of-mouth, internet/social media ads, etc.).

Question-2.2

What level of sales (with all values rounded to the nearest ten cents) would this fitted model predict for a company that allocates $3,000$ USD to newspaper advertisements?

Solution:

$$12,351.40 + (3) \cdot 54.70 = \boxed{12,515.50 \text{ USD}}$$

Question-2.3

Check all coefficients that are statistically significant at the $\alpha = 0.05$ significance level

- □ The `Intercept` coefficient
- □ The coefficient on `newspaper`

Solution: Both coefficients are statistically significant at the $\alpha = 0.05$ significance level, as we can infer from the fact that the values in the `P>|t|` column above are both less than `0.05`.

Question-2.4

Check all coefficients that are statistically significant at the $\alpha = 0.11$ significance level (the 89% "confidence" level[5]):

- □ The `Intercept` coefficient
- □ The coefficient on `newspaper`

Solution: Both coefficients are statistically significant at the $\alpha = 0.11$ significance level, as we can infer from the fact that the values in the `P>|t|` column above are both less than `0.11`.

Topic-3: Interpreting MLR Output

The following table shows the output from a **multiple** linear regression model, of `sales` on `TV`, `radio`, and `newspaper` (from the same `Advertising.csv` dataset), estimated using `statsmodels`.

```python
import pandas as pd
import statsmodels.formula.api as smf
ad_df = pd.read_csv("Advertising.csv").drop(columns='Unnamed: 0')
mlr_model = smf.ols('sales ~ TV + radio + newspaper', data=ad_df)
mlr_model_fitted = mlr_model.fit()
mlr_model_summary = mlr_model_fitted.summary(slim=True)
mlr_model_summary.extra_txt = ''
print(mlr_model_summary)
```

[5] Remember that, if someone asks you "why you care about the 89% confidence level?", you reply that it's because 89 is a prime number, which is a better answer than the answer to "why do you care about 95% confidence?", which is "because someone told me to use 95%, idk"... See Richard McElreath's *Statistical Rethinking* textbook for a full rant about this

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.897
Model:                            OLS   Adj. R-squared:                  0.896
No. Observations:                 200   F-statistic:                     570.3
Covariance Type:            nonrobust   Prob (F-statistic):           1.58e-
96
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.9389      0.312      9.422      0.000       2.324       3.554
TV             0.0458      0.001     32.809      0.000       0.043       0.049
radio          0.1885      0.009     21.893      0.000       0.172       0.206
newspaper     -0.0010      0.006     -0.177      0.860      -0.013       0.011
==============================================================================
```

Since we now have multiple independent variables in the same model, we also provide the correlation matrix among the variables in the regression here:

```
ad_df.corr()
```

```
                 TV     radio  newspaper     sales
TV         1.000000  0.054809   0.056648  0.782224
radio      0.054809  1.000000   0.354104  0.576223
newspaper  0.056648  0.354104   1.000000  0.228299
sales      0.782224  0.576223   0.228299  1.000000
```

Use this output to answer the following questions:

Question-3.1

What level of sales (with all values rounded to the nearest ten cents) would this fitted model predict for a company that allocates:

- \$1000 to TV advertisements,
- \$2000 to radio advertisements, and
- \$3000 to newspaper advertisements?

Solution:

$$2,938.90 + 1 \cdot 45.80 + 2 \cdot 188.50 + 3 \cdot (-1.00) = \boxed{3,358.70 \text{ USD}}$$

Question-3.2

Check all coefficients that are statistically significant at the $\alpha = 0.11$ significance level (89% "confidence" level)

- □ The `Intercept` coefficient
- □ The coefficient on `TV`
- □ The coefficient on `radio`
- □ The coefficient on `newspaper`

Solution: The coefficients on `Intercept`, `TV`, and `radio` are statistically significant at this significance level, while the coefficient on `newspaper` is not, since the p-value `0.860` shown in the table is well above the "cutoff" value of `0.11`.

Question-3.3

The p-value associated with the `newspaper` coefficient changed drastically between the SLR model in Problem-2 and the MLR model in this problem. Which of the following most likely explains this change in the regression results for the same independent variable?

(a) The p-values should be the same; there is an issue with the MLR result, resulting from the fact that we didn't tell `statsmodels` to estimate heteroskedasticity-robust standard errors
(b) The p-values should be the same; there is an issue with the MLR result, which would be fixed if we placed `newspaper` first among the list of independent variables in our call to `smf.ols()` in this part
(c) This difference between p-values is to be expected given the high correlation coefficient between `newspaper` and `radio` shown in the provided correlation matrix
(d) The p-values would converge to the same number if we increased our sample size more and more

Solution: (c) This change in p-values is to be expected given the high correlation coefficient between `newspaper` and `radio` shown in the provided correlation matrix.

This issue (of why we might obtain different SLR and MLR coefficients for the **same** independent variable) is discussed in-depth on pages 82-83 of ISLP!