STAT 151A Final Exam practice questions

Spring 2020

May 9, 2020

Instructions: answer each of the questions below. Please be sure to give clear reasons for your answers when asked and to show work to receive partial credit. You do not need to reduce all numerical calculations to a single number.

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

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(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
```

   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:

```
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)    XXXXX       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
```

(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**)

The sum of the dummy variables `x1-x4` *is equal to the intercept column in the design matrix, which means that the design matrix contains linearly dependent and is singular. It is not possible to uniquely define a least-squares solution for this design matrix because there is not a uniquely-defined inverse for $X^T X$. To solve the problem, one should eliminate one of the columns from the design matrix to ensure it is nonsingular. R appears to have done this automatically, elliminating* `x4`.

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

Since the dummy variable for seniors is left out of the regression and each of the other three groups has its own intercept term, the overall intercept is just the mean for seniors, which is 71.22. The residual standard error is a function of the ErrSS, which in turn is a weighted sum of the sample variances $\widehat{\sigma}_j^2$ for each of the four groups $j = 1, \ldots, 4$. The weights are the size of each group minus 1:

$$\sqrt{\widehat{\sigma}^2} = \sqrt{\frac{ErrSS}{n-p-1}} = \sqrt{\frac{19\sigma_1^2 + 19\sigma_2^2 + 19\sigma_3^2 + 19\sigma_4^2}{n-p-1}}$$
$$= \sqrt{\frac{19}{76}} \cdot \sqrt{3.92^2 + 3.91^2 + 6.92^2 + 5.77^2} \approx 5.29.$$

Finally, the adjusted R^2 is equal to

$$1 - \frac{n-1}{n-p-1} \cdot \frac{ErrSS}{TotSS} = 1 - \frac{n-1}{n-p-1}(1 - R^2) = 1 - \frac{79}{76}(1 - 0.4734) \approx 0.4526.$$

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

The standard error estimates for the regression coefficients are the square roots of the diagonal entries in the matrix $\widehat{\sigma}(X^TX)^{-1}$, where X is the design matrix that leaves out `x4`*. For this problem the value of X^TX is as follows:*

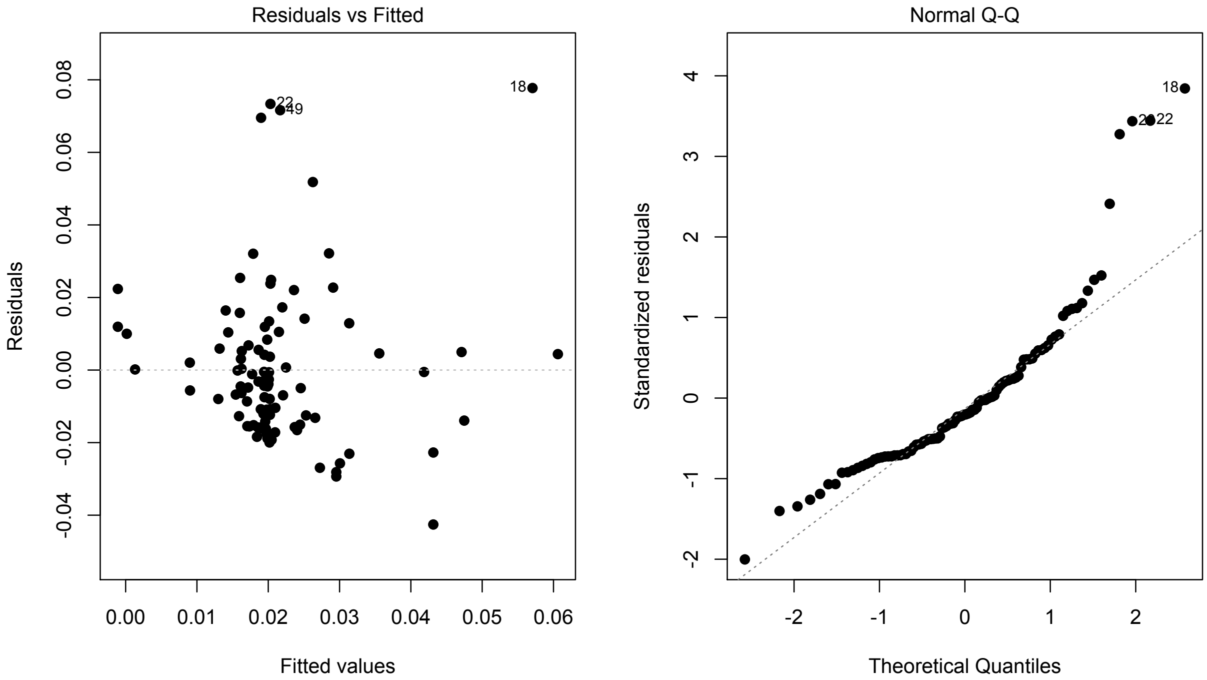
$$\begin{pmatrix} 80 & 20 & 20 & 20 \\ 20 & 20 & 0 & 0 \\ 20 & 0 & 20 & 0 \\ 20 & 0 & 0 & 20 \end{pmatrix}$$

This matrix (and hence its inverse) is unchanged if we shuffle the order of the categories, since the sample variances of `x1-x3` *and their sample covariances with each other and with the intercept are identical. So the diagonal values for* `x1-x3` *in matrix $\widehat{\sigma}(X^TX)^{-1}$ must also be identical.*

2. We run a linear regression with the response income and explanatory variables age and ethnicity. Ethnicity takes on four values: “White”, “Black”, “Hispanic”, “Asian”

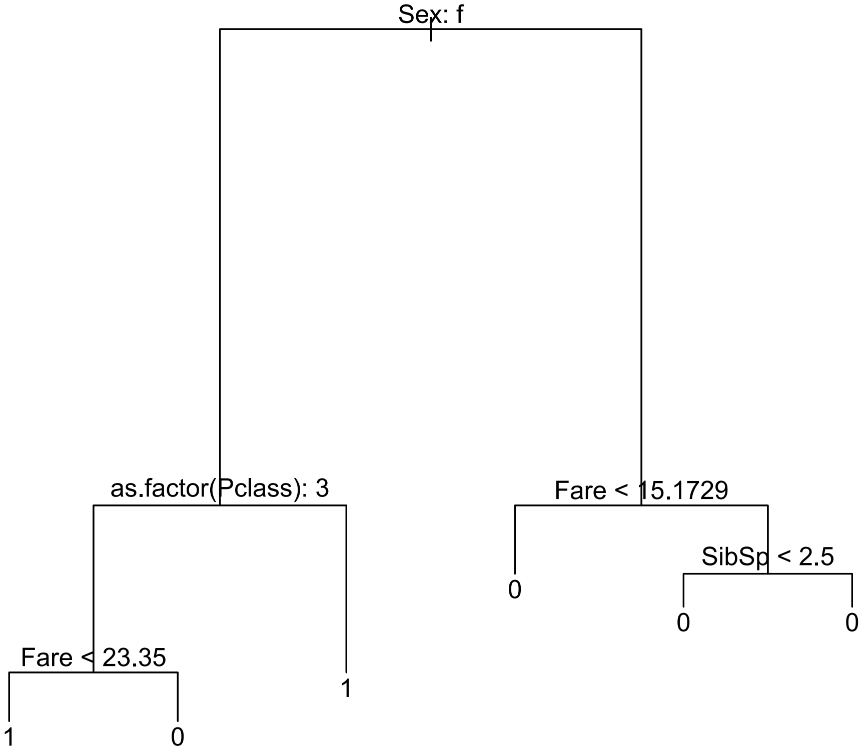
```
>  summar y(lm(y˜age*ethnicity))

Call:
lm(formula = y ˜ age * ethnicity)

Residuals:
     Min       1Q   Median       3Q      Max
-2.12271 -0.69124 -0.02662  0.61052  2.61588

Coefficients:
                        Estimate Std. Error   t value Pr(>|t|)
(Intercept)            5.000e+04  5.884e-01 84971.706  < 2e-16 ***
age                    4.028e-02  1.376e-02     2.928  0.00430 **
ethnicityBlack         8.936e-01  1.201e+00     0.744  0.45886
ethnicityHispanic      5.705e-01  8.090e-01     0.705  0.48246
ethnicityAsian        -1.612e+00  1.933e+00    -0.834  0.40649
age:ethnicityBlack    -7.941e-02  2.838e-02    -2.798  0.00625 **
age:ethnicityHispanic -7.880e-02  1.895e-02    -4.159 7.16e-05 ***
age:ethnicityAsian     1.176e-02  4.406e-02     0.267  0.79024
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.009 on 92 degrees of freedom
Multiple R-squared: 0.6596,Adjusted R-squared: 0.6337
F-statistic: 25.47 on 7 and 92 DF,  p-value: < 2.2e-16
```

Below are the residual vs. fitted plot and the Q-Q plot from the model. Describe what problems you see, if any, in the assumptions of the model. If you see problems in these diagnostic plots, describe what you might suggest to get an improved regression model. (**6 points**).



The main message I draw from the plots is that the residuals have a strong right skew. In particular, I notice that in both plots the largest positive residuals have a much greater magnitude than the largest negative residuals; this is especially clear in the QQ plot where we see that the right tail of the residual distribution is thicker than expected under the Normal distribution, while the left tail is thinner than expected. In the fitted vs. residual plot we also notice that there appears to be a linear boundary running from the origin through the point $(0.03, -0.04)$ *below which no points appear, creating an empty space in the bottom left of the plot; this is likely because income values never drop below zero in our dataset, so that observations with very small fitted values tend not to have negative residuals. To fix both these issues, I would try a log transformation on the response before fitting the model. The log transformation would spread out the small values of income relative to the large values and hopefully create a more symmetric distribution of residuals.*

3. We revisit a subset of the dataset *titanic* from the homework. The response variable is *Survived* which takes the value 1 if the passenger survived and 0 otherwise. We fit a classification tree to this dataset with the response variable being *Survived* and the explanatory variables being *Pclass* (proxy for the class in which the passenger travelled, with three levels 1, 2 and 3), *Sex* (gender), *SibSp* (number of siblings/spouses aboard), *Parch* (number of Parents/Children aboard), *Fare* (ticket fare) and *Embarked* (port of embarkation; has three levels: *C* for Cherbourg, *Q* for Queenstown and *S* for Southampton). The following plot summarizes the fitted tree.



(a) Based on the fitted tree, do we predict that a female passenger who travelled in *Pclass* 3 with a fare of 15 will survive or not? (**2 points**).

Yes, we expect that this individual will survive; the tree will sort her into the left-most leaf, for which the predicted value of survival is 1.

(b) Again based on the tree *rt*, do we predict that a female passenger who travelled in *Pclass* 1 with a fare of 25 will survive or not? (**2 points**).

Yes, we expect that this individual will survive; the tree will sort her into the third leaf from the left (since she is female and not in Pclass 3) for which the predicted value of survival is 1.

(c) The leaf of the tree oriented furthest to the right in the plot contains 186 individuals, 64 of whom survived. Calculate the Gini index at this leaf (**3 points**).

The Gini index for this node is given by

$$\sum_{i\in\{0,1\}} \overline{p}_i(1-\overline{p}_i)$$

where $\overline{p}_i$ *is the proportion of people in this node with survival status* i*. Substituting the sample proportions, we get*

$$\left(\frac{122}{186}\right)\left(1-\frac{122}{186}\right)+\left(\frac{64}{186}\right)\left(1-\frac{64}{186}\right)\approx 0.451.$$

If we were computing the Gini index for several nodes, we would multiply this Gini index by the total number of individuals at this leaf (186), which would give us 84.0.

(d) The decision rule SibSp < 2.5 at the far right of the plot separates two leaves, but both leaves give the same prediction. Assuming that the tree was grown using the Gini index, explain why this might have occurred (**4 points**).

The Gini index for a tree may be decreased by splitting a node even if both resulting leaves make the same prediction. For example, suppose that the second-to-rightmost leaf (with individuals for whom SibSp < 2.5) has 10 individuals, none of whom survived. The Gini index for this leaf is zero, so from the results of the previous question the total weighted Gini index across the two leaves is 84.0 If instead we combined the leaves, the weighted Gini index would be $196 \cdot 2(64/196)(132/196) \approx 86.2$, *which is larger.*

4. The summary of a fitted model is:

```
Coefficients:
            Estimate Std. Error t value
(Intercept)    1.12      1.00    1.12
x              0.50      0.25    2.0

Residual standard error: 2.2 on 40 degrees of freedom
Multiple R-squared:  0.22,
F-statistic: 2.27 on 1 and 40 DF
```

Below are the percentiles for a bootstrapped estimate of the sampling distribution of the studentized $\hat{\beta}$ in this simple linear model.

0.01	0.015	0.02	0.025	0.03	0.035	.04	0.045	0.05	0.055
-2.3	-2.0	-1.7	-1.6	-1.5	-1.4	-1.3	-1.3	-1.2	-1.2

0.95	0.955	0.95	0.965	0.97	0.975	0.98	0.985	0.99	0.995
1.7	1.7	1.7	1.9	2.0	2.0	2.2	2.4	2.7	3.4

Construct a 95% confidence interval for β, using the bootstrapped studentized distribution. (**6 points**).

The formula for the studentized bootstrap interval is

$$[\hat{\beta} - q^*_{(1-\alpha/2)}\hat{SE}(\hat{\beta}), \hat{\beta} - q^*_{(\alpha/2)}\hat{SE}(\hat{\beta})]$$

where the q_p *terms are quantiles of the bootstrapped studentized statistics. For a 95% interval, we set* $\alpha = 0.05$. *From the regression output we have* $\hat{\beta} = 0.5$ *and* $\hat{SE}(\hat{\beta}) = 0.25$. *From the percentile output above, we have* $q_{(1-\alpha/2)} = 2.0$. *and* $q_{(\alpha/2)} = -1.6$. *Putting it all together, the interval is*

$$[0.5 - 2 \cdot 0.25, 0.5 + 1.6 \cdot 0.25] = [0, 0.9].$$

5. We have data for 195 new high school students who each must choose one of three curricular tracks: academic, general, or vocational. We fit a multinomial logit model to predict their probabilities for choosing each option, using as explanatory variables a dummy variable for female gender and prior test scores for reading, writing, and math. Here is a summary of the fitted model:

```
Call:
multinom(formula = prog ˜ female + read + write + math, data = hschoice)

Coefficients:
         (Intercept) femalefemale     read     write    math
academic       -4.92        0.084  0.02860  0.000294  0.0781
vocation        4.10        0.242 -0.00888 -0.050700 -0.0246

Std. Errors:
```

```
         (Intercept) femalefemale   read  write   math
academic        1.43        0.413 0.0269 0.0295 0.0307
vocation        1.55        0.470 0.0305 0.0318 0.0355

Residual Deviance: 337.7572
AIC: XXXXXXXXX
```

(a) Fill in the missing value in the model output. (**2 points**).

AIC is defined as $-2\log\mathcal{L}+2k$ *where* $\mathcal{L}$ *is the likelihood for the model and* k *is the number of parameters in the model, which in this case is 10 (5 coefficients each for modeling academic and vocational tracks). In addition, the residual deviance is defined as* $-2(\log\mathcal{L}-\log\mathcal{L}_{sat})$ *where* $\mathcal{L}_{sat}$ *is the likelihood of the saturated model. Since* $\mathcal{L}_{sat}=1$*, the residual deviance is just* $-2\log\mathcal{L}$ *and*

$$\textit{AIC} = \textit{Residual Deviance } + 2(p+1) = 337.7572 + 2(10) = 357.7572.$$

(b) Rank the predicted probabilities for academic, general, and vocational tracks from largest to smallest for a male student with scores of 40 for reading and writing and a score of 30 for math. (**4 points**). *The multinomial logit model tells us*

$$\log\left(\frac{\pi_{acad}}{\pi_{gen}}\right) = X\beta_{acad} \qquad \log\left(\frac{\pi_{voc}}{\pi_{gen}}\right) = X\beta_{voc} \qquad \log\left(\frac{\pi_{gen}}{\pi_{gen}}\right) = 0$$

If we add $\log(\pi_{gen})$ *to each quantity above and exponentiate the result, we'll get the raw probabilities* π_{acad}, π_{voc}*, and* π_{gen}*; however, since both those transformations are monotonic, the rankings of these quantities will be the same as the ranking of the probabilities. For the individual described, the linear combinations* $X\beta_{acad}$ *and* $X\beta_{voc}$ *are equal to* $-4.92+(0.02860+0.000294)40+(0.0781)30 \approx -1.42$ *and* $4.10+(-0.0088-0.0507)40+(-0.0246)30 \approx 1.65$*. Therefore since* $1.65 > 0 > -1.42$*, the individual is most likely to take the vocational track, followed by the general track, followed by the academic track.*

(c) Suppose we want to conduct a hypothesis test for whether the dummy variable for female gender is necessary in this model. Write the associated null hypothesis in mathematical notation, describe a test statistic we could use to test it, and give its null distribution. (**4 points**).

Let $\gamma_{acad,female}$ *and* $\gamma_{voc,female}$ *be the two regression coefficients associated with female gender in the model. The null hypothesis we wish to test is* $H_0 : \gamma_{voc,female} = \gamma_{voc,female} = 0$*. Let the likelihood of the original model be* $\mathcal{L}_1$ *and the likelihood of the reduced model omitting the dummy variable for female gender be given by* $\mathcal{L}_0$*. We can test* H_0 *using the likelihood ratio statistic* $-2\log(\mathcal{L}_0/\mathcal{L}_1)$*, which will have a* χ^2_2 *distribution since the models differ by two degrees of freedom (for the two parameters set to zero under* H_0*).*

(d) Now suppose we combine the students choosing academic and general tracks and instead model the choice of vocational or non-vocational tracks, using logistic regression with all four explanatory variables in the original model. Compute and the degrees of freedom for this logistic model and compare it to the degrees of freedom for the original multinomial logistic model. (**2 points**).

The logistic regression model contains one degree of freedom for each explanatory variable plus one more for the intercept, for a total of five degrees of freedom. The multinomial logistic model has five degrees of freedom for modeling the odds of the academic track relative to the general track, and another five degrees of freedom for modeling the odds of the vocational track relative to the general track, for a total of 10. By combining two of our three categories, we cut our degrees of freedom in half.

6. Using a subset of the `Boston` data from class, we model median home value (`medv`) as a nonlinear function of the percentage of low-socieconomic-status residents in each census tract (`lstat`). We use cubic polynomial regression, with an intercept and regressors `lstat`, `lstat`^{2}, and `lstat`^{3}. The fitted coefficients for the different basis elements are as follows:

Basis element	Estimate	Std. Error
(Intercept)	48.93	1.53
`lstat`	-3.733	0.349
`lstat`^{2}	0.1419	0.0226
`lstat`^{3}	-0.001915	0.000427

Residual standard error: 5.397 on 446 degrees of freedom

(a) Compute the fitted value for median home value for a census tract with 10 percent low-socioeconomic status residents. (**4 points**).

The fitted value is

$$48.93 - 3.733 \cdot 10 + 0.1419 \cdot 100 - 0.001915 \cdot 1000 = 23.875.$$

(b) Write down an expression giving an approximate 95% confidence interval for the fitted value in the previous part, as a function of $(X^TX)^{-1}$,where X is the design matrix for the polynomial basis (including the intercept term). Please substitute numbers for the other parts of the expression. (**6 points**).

The 95% confidence interval is given by

$$\widehat{\beta} \pm 1.96 \cdot \widehat{\sigma}\sqrt{x_i(X^TX)^{-1}x_i^T} = 23.875 \pm 1.96 \cdot 5.397 \sqrt{\begin{pmatrix} 1 \\ 10 \\ 100 \\ 1000 \end{pmatrix} (X^TX)^{-1} \begin{pmatrix} 1 & 10 & 100 & 1000 \end{pmatrix}}.$$

(c) Suppose now we fit a cubic regression spline instead of a polynomial, introducing one knot at the median value of `lstat`. How should we change the design matrix to fit this model instead? How many degrees of freedom will the model now use? (**4 points**).

We should change the design matrix by adding one column equal to $(\texttt{lstat} - c)_+^3$ where c is the median value of `lstat`. This will expend one more degree of freedom, so the model now uses a total of 5 instead of 4.

7. Answer TRUE or FALSE to the following statements **and justify your answer**.

(a) The leverage for the ith subject measures how far the ith subject is from the rest of the subjects in terms of the explanatory variable values. (**2 points**).

TRUE. The leverage for the ith observation is given by the diagonal entry of the hat matrix, which can be interpreted as a weighted distance between observation i and the vector of means for the regressors $(\overline{x}_1, \ldots, \overline{x}_p)$ in the vector space defined by the columns of the design matrix.

(b) The MLE of β in a logistic regression model can always be computed in closed form. (**2 points**).

FALSE. The MLE for logistic regression is a solution to a nonlinear system of equations that has no closed form in general.

(c) Transformations of the explanatory variables can potentially improve the fit of the model in logistic regression. (**2 points**).

TRUE. For example, suppose the probability of success appears to have a U-shaped relationship with an explanatory variable, where the probability of success is high for values of x with large magnitude, whether positive or negative, but is low for values of x near zero. In this case the transformation x^2 will likely improve the model's fit.

(d) In a model with p continuous explanatory variables, Mallow's C_p must be $\geq p + 1$. (**2 points**).

FALSE. The formula for Cp is $p + 1 + (k - p)(F - 1)$ where k is the number of non-intercept parameters in the full model, p is the number in the model under current consideration, and F is the F-statistic comparing the reduced and full. models. Cp will be lower than $p + 1$ whenever $p < k$ and $F < 1$ (an event which happens with nonzero probability under the null hypothesis that the reduced model is correct).

(e) R^2 is an effective model selection criterion for deciding the best size for a linear model. (**2 points**).

FALSE. R^2 always increases as more variables are added to the linear model, so it will always choose a larger model over a submodel, even if the additional covariates used are pure noise.

(f) In simple regression, rejecting the null hypothesis that $\beta = 0$ is equivalent to concluding that x has a causal effect on y. (**2 points**).

FALSE. Even if changing x has no effect on y, if both x and y are highly correlated with a third variable z not present in the regression, then x and y may have a strong correlation. In this case, the null hypothesis $\beta = 0$ will likely be rejected but this is not enough to demonstrate a causal effect.