Final Exam - DSC 152, Spring 2026

Full Name:	SOLUTIONS
PID:	

Lecture Section: ◯ 8:00am ◯ 9:30am

Instructions:

- This final exam consists of 5 questions for a total of 116 points. You have a total of 3 hours to complete it.
- Please write **clearly** in the provided answer boxes; we will not grade work that appears elsewhere. Completely fill in bubbles and square boxes; if we cannot tell which option(s) you selected, you may lose points.
  - ◯ A bubble means that you should only **select one choice**.
  - ☐ A square box means you should **select all that apply**.
- You may use one handwritten sheet of notes. No calculators and no computers.
- Assume we have already run all necessary `library()` calls in R.

By signing below, you are agreeing that you will behave honestly and fairly during and after this exam.

Signature:

Version A

Please do not open your exam until instructed to do so.

You are analyzing a dataset of San Diego public schools called `sd_schools`. Each row corresponds to one school and includes the following variables:

- `school_name` (character)
- `type` (character): Public Charter or Public District
- `level` (character): Elementary / Middle / High / K-8 / Alternative
- `gs` (double): Rating of the school according to the GreatSchools website, 1–10 (higher is better)
- `enroll` (double): Total enrollment
- `s_t_r` (double): Student-to-teacher ratio
- `pct_lcff` (double): Percentage of families in the school who financially qualify for need-based assistance
- `ela` (double): Percentage of the school meeting statewide English Language Arts (ELA) standards
- `math` (double): Percentage of the school meeting statewide math standards

A random sample of rows are shown below.

school_name	type	level	gs	enroll	s_t_r	pct_lcff	ela	math
Adams Elementary	Public District	Elementary	4	330	21.6	62	32	24
B. Pendleton Elem.	Public District	Elementary	3	300	21.1	86	14	9
Miller Elementary	Public District	Elementary	6	400	22.0	38	48	40
Clairemont Canyons Acad.	Public Charter	Elementary	6	520	22.5	40	48	40
Magnolia Science Acad.	Public Charter	Middle	5	450	19.5	55	38	30
O'Farrell Community	Public Charter	Middle	5	650	22.0	70	34	24

PID: _____

Question 1

A researcher is interested in inference on whether schools are above a 30% threshold for math preparation, with student-to-teacher ratio as the covariate of interest. She creates a corresponding binary outcome variable and fits a logistic regression:

```
sd_schools$math_satisfactory <- ifelse(sd_schools$math >= 30, 1, 0)

model1 <- glm(math_satisfactory ~ s_t_r, data = sd_schools,
                                  family = "binomial")
```

The coefficients table from `summary(model1)$coefficients` is:

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	−14.32	3.97	−3.61	0.0003
`s_t_r`	0.6585	0.183	3.60	0.0003

a) (5 pts) Which of the following are true in general about logistic regression? **Select all that apply.**

■ The outcome variable must be binary (takes values 0 or 1).

☐ The covariate(s) must be binary (takes values 0 or 1).

☐ The coefficients are estimated by minimizing the RSS to fit a linear relationship between the logit-transformed values of the binary outcome and the predictors.

■ The fitted values $\hat{y}_i$ from a logistic regression model are always between 0 and 1.

☐ Since logistic regression gives estimates of odds ratios, it is only valid for case-control data.

☐ None of the above.

b) (5 pts) Using the coefficients table above, write the fitted model equation for the log-odds of `math_satisfactory`. Define any notation you introduce.

Solution: Let $\hat{y}_i$ be the predicted value of `math_satisfactory` at $\texttt{s\_t\_r}_i$. Then:

$$\log\left(\frac{\hat{y}_i}{1-\hat{y}_i}\right) = -14.32 + 0.6585 \cdot \texttt{s\_t\_r}_i.$$

c) (4 pts) The "crossover point" is the value of `s_t_r` at which the predicted probability equals exactly 0.5. Write a general expression for this crossover value in terms of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, then evaluate it for `model1`. You may leave your answer as a fraction. *Hint: it may or may not be helpful to note that log(1) = 0.*

Solution: Setting the log-odds equal to 0: $0 = \widehat{\beta}_0 + \widehat{\beta}_1 x^*$, so $x^* = -\widehat{\beta}_0/\widehat{\beta}_1$. For `model1`:

$$x^* = -\frac{-14.32}{0.6585} = \frac{14.32}{0.6585} \approx 21.75.$$

d) (3 pts) What are the three conditions required for validity of logistic regression inference, according to what we learned in this class? List them.

Solution:

1. The outcome variable is binary.
2. Observations are independent.
3. The relationship between each predictor and the log-odds of the outcome is linear.

e) (5 pts) The p-value for each coefficient is used to test $H_0 \colon \beta_j = 0$. If $\alpha = 0.05$, which of the following are true? **Select all that apply.**

☐ A Type I error occurs when we conclude that `s_t_r` is not a significant predictor when it in fact has no effect in the population.

☐ A very large sample size guarantees that any statistically significant result is always practically meaningful.

■ If the true coefficient for `s_t_r` is zero, a Type I error has been made here.

☐ Reducing α from 0.05 to 0.01 would reduce our Type I error rate and increase our power.

■ The p-value for `(Intercept)` tells us whether the odds of meeting math proficiency is different from 1 for a student-teacher ratio of 0, according to the fitted model.

☐ None of the above.

PID: _____

Question 2

Using the same `math_satisfactory` variable, researchers ask: are charter schools more or less likely to be math-satisfactory compared to public district schools? Treat Public Charter as the "exposed" group, Public District as the "unexposed" group, and `math_satisfactory` as the outcome of interest. The cross-tabulation of the data gives:

	Outcome = 1	**Outcome = 0**	**Total**
Exposed (Charter)	$a = 8$	$b = 3$	11
Unexposed (District)	$c = 83$	$d = 85$	168
Totals	91	88	179

a) (2 pts) Using the numeric values corresponding to a, b, c, d from the table above, write down the expressions for the relative risk (RR) and the odds ratio (OR) as unsimplified fractions.

Solution:

$$\text{RR} = \frac{a/(a+b)}{c/(c+d)} = \frac{8/11}{83/168} \qquad \text{OR} = \frac{ad}{bc} = \frac{8 \cdot 85}{3 \cdot 83}$$

b) (2 pts) The OR from part (a) is larger than the RR, and both are above 1. Which one of the following best describes why? Select one.

- ◯ The sample sizes are different in the two groups.
- ◯ The RR is always a better measure of association than the OR.
- ◯ The OR always overestimates the RR when the exposure is present.
- ● The OR is only a good estimate of the RR when the outcome is rare.
- ◯ None of the above.

c) (10 pts) Write R code to carry out a permutation test for the OR, testing the **one-sided** alternative that the OR is greater than 1 (i.e., charter schools have higher odds of being math-satisfactory).

Your code may assume that column 2 of `sd_schools` is the school `type` and column 10 is `math_satisfactory`, and these column numbers may be hardcoded in your answers below. Fill in the boxes with the required code corresponding to each letter.

```
calc_OR <- function(df){
  a <- _______(a)________
  b <- _______(b)________
  c <- _______(c)________
  d <- _______(d)________
  return( ______(e)_______ )
}

null_OR <- NULL
reps <- 10000
for(i in 1:reps){
  perm_df <- sd_schools
  perm_df[,2] <- _______(f)_______
  null_OR[i] <- calc_OR(perm_df)
}

sum( _________(g)___________ ) / reps  # calculate p-value
```

(a): `sum(df[,2]=="Public Charter" & df[,10]==1)`

(b): `sum(df[,2]=="Public Charter" & df[,10]==0)`

(c): `sum(df[,2]=="Public District" & df[,10]==1)`

(d): `sum(df[,2]=="Public District" & df[,10]==0)`

(e): `(a*d) / (b*c)`

(f): `sample(perm_df[,2])`

(g): `null_OR >= calc_OR(sd_schools)`

d) (4 pts) The 95% confidence interval for the odds ratio from the 2×2 table is (0.76, 12.79). The following logistic regression is also fit:

```
model2 <- glm(math_satisfactory ~ school_type, data = sd_schools,
              family = "binomial")
```

with the following output:

```
summary(model2)$coefficients
```

```
##                               Estimate  Std. Error z value Pr(>|z|)
## (Intercept)                  0.9808293 0.6770027  1.448782 0.1473985
## school_typePublic District -1.0046399 0.6943670 -1.446843 0.1479410
```

and 95% CI:

```
exp(confint(model2))[2,]
```

```
## 2.5 % 97.5 %
## 0.07817285 1.31450289
```

Which of the following are true? **Select all that apply.**

☐ Logistic regression is not an appropriate way to obtain an odds ratio for 2×2 data, so it is not surprising that these results would differ.

☐ The confidence interval from `model2` is for the log odds ratio, so it is not surprising that it differs from the CI for the OR from the 2×2 table.

■ Since the baseline level in `model2` is "Public Charter," this is backwards from the 2×2 table OR, so the CI from `model2` is simply the reciprocal of the one from the table.

■ The result from `model2` indicates that H_0: OR $= 1$ would not be rejected using the data in the 2×2 table.

☐ None of the above.

e) (5 pts) As a sensitivity analysis, a researcher adds `school_type` as a covariate in `model1` and fits:

```
model3 <- glm(math_satisfactory ~ s_t_r + school_type,
                                  data = sd_schools,
                                  family = "binomial")
summary(model3)$coefficients
```

```
                              Estimate Std. Error   z value Pr(>|z|)
(Intercept)                -23.414301  4.5819028 -5.110170 0.00000032
s_t_r                        1.295187  0.2428253  5.333822 0.00000009
school_typePublic District  -5.024487  1.2020425 -4.179958 0.00002915
```

Based on this and all preceding parts, which of the following are correct? **Select all that apply.**

■ `s_t_r` appears to have a stronger impact on `math_satisfactory` when adjusting for `school_type` than in the unadjusted `model1`.

■ This output shows evidence that, if analyses are performed on another set of newly obtained schools, `school_type` should be included in the model for inference.

☐ At $\alpha = 0.05$, `school_type` is statistically significant in `model3` but not in `model1`, so one of these analyses must be incorrect.

☐ From the `model3` output, $e^{1.2952}$ is the estimated OR for meeting math proficiency corresponding to a 1-unit increase in `s_t_r`, but only among "Public District" schools.

☐ From the `model3` output, $e^{1.2952}$ is the estimated OR for meeting math proficiency corresponding to a 1-unit increase in `s_t_r`, but only among "Public Charter" schools.

☐ None of the above.

f) (4 pts) The researcher wants to print the estimated odds ratio for `school_type` directly from `model3`. Write one line of R code to produce **only** the estimated odds ratio for `school_type`.

Solution: `exp(summary(model3)$coefficients[3, 1])`

PID: _____

Question 3

Researchers fit the following model:

```
full_mod <- lm(gs ~ pct_lcff + s_t_r + ela + math + level,
                          data = sd_schools)
summary(full_mod)
```

The output is shown below.

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.363560   0.557111   0.653  0.51491
pct_lcff               0.009369   0.003637   2.576  0.01085 *
s_t_r                 -0.038609   0.017965  -2.149  0.03304 *
ela                    0.140174   0.018289   7.665 1.31e-12 ***
math                  -0.031854   0.017312  -1.840  0.06752 .
levelElementary        0.525420   0.201368   2.609  0.00988 **
levelHigh              0.447701   0.218581   2.048  0.04208 *
levelK-8               0.516680   0.210668   2.453  0.01519 *
levelMiddle            0.622939   0.211081   2.951  0.00361 **

Residual standard error: 0.2769 on 170 degrees of freedom
Multiple R-squared:  0.9789,  Adjusted R-squared:  0.978
F-statistic: 988.2 on 8 and 170 DF,  p-value: < 2.2e-16
```

a) (3 pts) Interpret the coefficient of `ela` in context.

> **Solution:** Holding all other predictors fixed, a 1 percentage-point increase in `ela` is associated with an estimated 0.1402-point increase in `gs`, on average.

b) (3 pts) Interpret the coefficient of `levelMiddle` in context.

> **Solution:** Holding the four quantitative predictors fixed, middle schools are estimated to have `gs` 0.6229 points higher than the reference level (Alternative schools), on average.

c) (2 pts) A student says: "Since `pct_lcff` has a positive coefficient in this model, schools with higher poverty tend to have higher GreatSchools ratings in the sample overall." Explain why this statement might not actually be true. *Hint: recall Simpson's Paradox.*

Solution: The coefficient is adjusted for the other variables in the model and does not describe the marginal relationship between `pct_lcff` and `gs`. The positive sign reflects the partial association after accounting for test scores and school level, not the overall trend.

d) (4 pts) Researchers want to test whether school `level` is associated with `gs`, after adjusting for the four quantitative predictors. State the null and alternative hypotheses in terms of regression parameters.

Solution: Let $\beta_E, \beta_H, \beta_K, \beta_M$ be the coefficients for Elementary, High, K-8, and Middle (reference = Alternative). Then:

$$H_0\colon \beta_E = \beta_H = \beta_K = \beta_M = 0 \quad \text{vs.} \quad H_A\colon \text{at least one is nonzero.}$$

e) (6 pts) Write R code to carry out the appropriate partial F-test for part (d).

Solution:
```
reduced_mod <- lm(gs ~ pct_lcff +
                  s_t_r + caaspp_ela_pct_met +
                  caaspp_math_pct_met, data = sd_schools)
anova(reduced_mod, full_mod)
```

f) (2 pts) Suppose the partial F-test gives $F = 2.54$, $p = 0.044$. At $\alpha = 0.05$, state the conclusion in context.

Solution: Since $p = 0.044 < 0.05$, we reject H_0. There is statistically significant evidence that school level is associated with `gs`, after adjusting for the four quantitative predictors.

g) (5 pts) Which of the following are true? **Select all that apply.**

☐ The p-value for `levelMiddle` alone answers whether `level` as a whole is associated with `gs`, after adjusting for the four quantitative predictors.

■ The partial F-test in part (e) works by comparing the RSS of a reduced model without `level` to the RSS of the full model with `level`.

■ Because `level` is coded using several indicator variables, testing whether `level` matters requires testing all of its corresponding coefficients together.

☐ The reduced model for this partial F-test should keep only the `level` coefficient with the smallest individual p-value and remove the other `level` coefficients.

☐ The global F-test from `summary(full_mod)` tests the same null hypothesis as the partial F-test for `level`.

☐ None of the above.

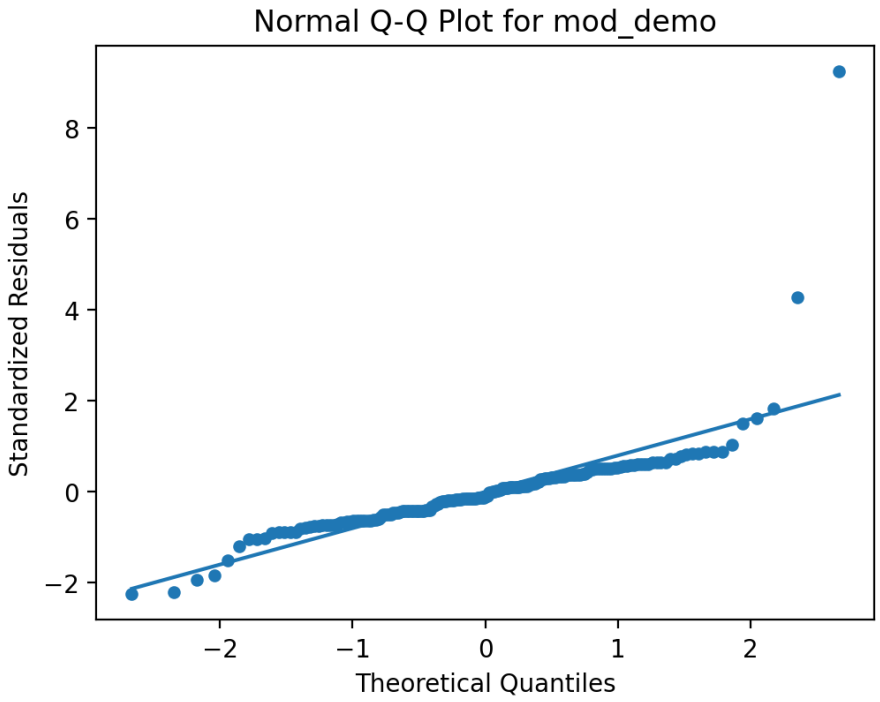
Question 4

Researchers fit the following model:

```
mod_demo <- lm(gs ~ school_type + pct_lcff + s_t_r,
                            data = sd_schools)
summary(mod_demo)$coefficients
```

```
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   9.331750   0.682605  13.671  < 2e-16 ***
school_typePublic District -1.285179   0.192710  -6.669 3.25e-10 ***
pct_lcff                   -0.071215   0.001928 -36.940  < 2e-16 ***
s_t_r                       0.034170   0.031645   1.080    0.282
```

The researchers produce the following Q-Q plot for the residuals from `mod_demo`:



a) (2 pts) What condition does the Q-Q plot primarily assess? Based on this plot, does the condition appear to hold? Briefly explain.

> **Solution:** The Q-Q plot assesses the **Normality** condition for the residuals. Here it appears to be **violated**: most residuals fall close to the reference line, but the upper tail deviates substantially (standardized residuals near 4 and 9), indicating a heavier right tail than expected under normality.

b) (2 pts) The coefficient for `school_typePublic District` is statistically significant. Explain why this does not, by itself, imply that `school_type` causes differences in `gs`.

Solution: This is observational data. School type is likely associated with other school and neighborhood characteristics that also affect ratings, so confounding may explain the observed association. Statistical significance does not imply causation.

c) (2 pts) Notice that the coefficient of `school_typePublic District` is negative in the output above. Which of the following is the best interpretation? Select one.

◯ Public district schools have lower `gs` than public charter schools on average, because `school_type` causes differences in `gs`.

● Among schools with the same `pct_lcff` and `s_t_r`, public district schools have lower predicted `gs` than public charter schools.

◯ Public district schools have lower predicted `gs` than public charter schools after adjusting for **all** possible confounders.

◯ Since `school_typePublic District` is statistically significant, `school_type` must be the most important predictor in the model.

◯ For every fixed value of `pct_lcff` and `s_t_r`, every public district school is predicted to have a lower `gs` than every public charter school.

d) (9 pts) Suppose the researchers are actually focused on the student-to-teacher ratio and are interested in whether the regression t-test for $H_0 \colon \beta_{\text{s\_t\_r}} = 0$ has an approximately correct Type I error rate when H_0 is true, given the observed associations for the other covariates with the outcome variable. Complete the R code below to run simulations under a null model where `s_t_r` has no effect, then estimate the actual Type I error rate when $\alpha = 0.05$.

Specifically:

- Assume that $\epsilon_i \sim N(\mu = 0, \sigma = 10)$
- `gs_sim` in the code below should be generated according to the fitted regression equation that can be gathered from the model output, with the exception of $\beta_{s\_t\_r}$ which is equal to 0 under H_0, and along with ϵ_i as noted in the previous bullet.

Fill in the required blanks in the code on the next page.

```
reps <- 1000
count <- 0

for (i in 1:reps) {
    # Generated simulate values of gs under H0
    gs_sim <- _____________(a)_____________

    sim_data <- sd_schools
    sim_data$gs <- gs_sim

    # Run the linear model that will give the necessary p-value
    sim_mod <- ____________(b)_____________

    # Get that p-value from the model output
    pval <- ____________(c)_____________
    if(pval < 0.05){
      count <- count + 1
    }
}

count/reps
```

Write the required code for blank (a) below:

Solution:

```
9.331750 - 1.285179*(sd_schools$type == "Public District")
 - 0.071215*sd_schools$pct_lcff + rnorm(dim(casino)[1], 0, 10)
```

Write the required code for blank (b) below:

Solution:

```
lm(gs ~ school_type + pct_lcff + s_t_r, data = sim_data)
```

Write the required code for blank (c) below:

Solution:

```
summary(sim_mod)$coefficients[4,4]
```

e) (2 pts) Suppose the simulation returns `count/reps = 0.087`. Interpret this number in context.

Solution: Across 1000 simulated datasets generated under a null model where school type has no effect, the t-test rejected H_0: $\beta_{\text{District}} = 0$ in about 8.7% of simulations.

f) (5 pts) Which of the following are valid interpretations of `count/reps = 0.087`? **Select all that apply.**

■ The estimated Type I error rate is about 8.7% for this simulation design.

■ With a nominal significance level of 5%, this simulation suggests the test may reject too often under the null.

☐ The probability that the null hypothesis is true is 8.7%.

☐ The power of the test is 8.7%.

☐ The simulation proves that `school_type` has no causal effect on `gs`.

☐ None of the above.

Question 5

For this question, we will move away from the San Diego public schools dataset and consider the data from the Lucas 2013 paper as discussed in class and on Lab 9. The dataframe is called `casino` and a snippet of some of the columns in it is here:

date	day_num	dow	R1_COIN	R1_DROP	R1_RAKE	R2_COIN
2009-02-03	1	Tuesday	1772621	276443.8	3971.705	5224949
2009-02-04	2	Wednesday	1772529	281302.1	4342.208	4883472
2009-02-05	3	Thursday	1880721	376328.5	4358.191	6125977
2009-02-06	4	Friday	2237850	544662.9	5653.995	8304579
2009-02-07	5	Saturday	2545841	678492.9	9668.532	9803198
2009-02-08	6	Sunday	2046482	394702.7	5527.945	7452677
2009-02-09	7	Monday	1776765	274851.6	5544.414	5340779
2009-02-10	8	Tuesday	1894535	280856.5	5153.511	6072844

These columns are:

- `date` (character): Calendar date in YEAR-MO-DA format
- `day_num` (double): Day number from 1 to 217
- `dow` (character): Day of the week
- `R1_COIN` (double): A measure of slot machine revenue in Resort 1 of the dataset
- `R1_DROP` (double): A measure of table game revenue in Resort 1 of the dataset
- `R1_RAKE` (double): A measure of poker room revenue in Resort 1 of the dataset
- `R2_COIN` (double): A measure of slot machine revenue in Resort 2 of the dataset

PID: _____

a) (5 pts) Consider an investigation of the relationship between `R1_RAKE` and `R1_COIN`: that is, poker room revenue and slot machine wagers respectively in Resort 1. An MA(2) model, ignoring any other possible covariates, is fit below:

	Estimate	Std. Error	z value	Pr(> |z|)
ma1	0.715	0.074	9.680	3.653928e-22
ma2	0.282	0.064	5.030	4.896851e-07
intercept	1288558.820	76712.064	16.797	2.552256e-63
xreg	135.699	11.191	12.126	7.714198e-34

Write the regression equation indicated by the coefficient estimates in the table above.

Solution:
$R1\_COIN_t = 1288558.82 + 135.699 \cdot R1\_RAKE_t + \epsilon_t + 0.715\epsilon_{t-1} + 0.282\epsilon_{t-2}$

b) (8 pts) Now, suppose that a researcher wanted to ignore the time-series nature of the data, but the data truly come from an MA(2) model. That is, the researcher wants to simply fit:

```
lm(R1_COIN ~ R1_RAKE, data=casino)
```

You would like to estimate via simulation what the resulting power of this analysis would be. Using the output from part (a) as the true coefficient values, write a simulation to estimate the power of the simple linear regression model if the data truly follow an MA(2) model with coefficient values from the output table. A few notes:

- For this simulation, assume that $\epsilon_t \sim N(\mu = 0, \sigma = 50000)$
- Recall that the full dataset has a total sample size of $n = 217$ as needed.
- The `R1_RAKE` values should be the same values that are in the `casino` dataframe as is; that is, you should use `casino$R1_RAKE` directly in your simulation as needed.

Here is some preliminary code:

```
beta0 <- 1288558.820
beta1 <- 135.699
theta1 <- 0.715
theta2 <- 0.282
count <- 0
reps <- 10000
```

Now continue the required code in the box on the next page.

Solution:

```
for(j in 1:reps){
 eps <- rnorm(219, sd=50000)
 y <- NULL

 for(i in 1:217){
 y[i] <- beta0 + beta1 * casino$R1_RAKE[i] + eps[i+2] + 0.715*eps[i+1]
+ 0.322*eps[i]
 }

 pval <- summary(lm(y ~ casino$R1_RAKE))$coefficients[2,4]

 if(pval < 0.05)
 count <- count + 1

}

count / reps
```

c) (6 pts) Now, consider the following output with `R1_COIN` still as the outcome variable, and with additional covariates:

	Estimate	Std. Error	z value	Pr(> |z|)
`ma1`	0.712	0.068	10.456	0.000
`ma2`	0.402	0.062	6.508	0.000
`intercept`	1582655.069	79003.511	20.033	0.000
`R1_RAKE`	72.795	13.049	5.578	0.000
`Friday`	274303.200	44596.169	6.151	0.000
`Saturday`	383101.482	56227.231	6.813	0.000
`JUN30`	1782.234	176016.966	0.010	0.992

Write R code below that would produce this output.

Solution:

```
X <- cbind(casino$R1_RAKE,
ifelse(casino$dow == "Friday", 1, 0),
ifelse(casino$dow == "Saturday", 1, 0),
ifelse(casino$date == "2009-06-30", 1, 0)
)

colnames(X) <- c("R1_RAKE", "Friday", "Saturday", "JUN30")

mod <- Arima(y=casino$R1_COIN, order=c(0,0,2), xreg=X)
tmp <- coeftest(mod)
```

d) (5 pts) In the Lucas 2013 paper, which of the following are shortcomings of the statistical analyses, according to what we have learned in this class? **Select all that apply.**

■ The models used for statistical inference were formulated via a model selection procedure on the same data that were used for inference.

☐ The data did not meet the independence condition required for linear regression.

☐ The author used a time series analysis when standard linear regression would have sufficed.

☐ The author's scientific question was ill-posed and does not translate to a valid statistical question

☐ The author extrapolated his conclusions to the entire population of all casinos when his data only came from three casinos in Las Vegas.

☐ None of the above.

Reminders:

- **Write your PID** on the front page and on the top right corner of each subsequent page.
- Fill in bubbles and square boxes **completely and darkly**; partially filled marks will not be graded.
- Show all written work and code clearly inside the response boxes; work outside boxes will not be graded.