STAT 207 STUDY GUIDE FOR EXAM 2 SOLUTIONS

The Exam 2 notebook will be released on the class git release repository as exam_02 by Monday, April 13, 9 am US central time.

Your completed Exam 2 notebook must be committed and pushed to your remote class repo by Tuesday, April 14, 11:59 pm US central time.

- The exam will be in the form of a Jupyter notebook. this is an open notes, open internet exam, **but you must do all the work independently without consulting other people.**
- The instructional staff will not answer questions about how to do the problems. They will only answer questions about possible typos or other errata in the exam.
- The exam will be intended to be a one hour exam if done in one sitting.
- The exam covers classnote chapters 8 - 12 and homework labs 6 - 9.

Exam problems are generally be shorter than homework problems and may involve short answer conceptual questions, quick calculations and Python code interpretation or debugging. It is not a multiple choice exam, although some multiple choice questions are possible.

Here is a list of concepts covered since the previous exam:

- Hypothesis testing: formulating hypotheses as statements about population parameters
- Hypothesis testing: large sample z tests for means and proportions
- Hypothesis testing: t distribution and degrees of freedom
- Hypothesis testing: one- and two-sample t tests and confidence intervals for means
- Scatter plots and correlation
- Simple linear regression: Relation between regression slope, correlation, S_y and S_x
- OLS model fitting criterion
- Model checking with residual plots
- Multiple regression models: interpretation of coefficents
- Multiple regression models: model summary, estimates, standard errors, coefficient t tests
- F test for the regression
- Oneway anova model
- Analysis of variance F tests: comparing two models
- Categorical data: Odds, odds-ratios, log-odds ratios, standard errors
- Logit models for binary responses (logistic regression models)
- Coefficent estimates, interpretation, standard errors, z-test, confidence intervals
- Converting the estimated log-odds to estimated probablity and vice versa
- Logit classifier
- Accuracy of classifier
- Sensitivy, Specificity, True Negative, False Positive, False Negative, True Positive

Below are some exam type practice problems. Note: Please do not expect the exam to be exactly like these sample questions. Use them as a way to see what you know or need to study further. The exam itself will not be as long as this list of practice problems.

1. Internal body temperature varies throughout the day. Suppose we obtain a random sample of 16 individuals who have their temperature taken immediately after waking. A 95% confidence interval is calculated for the true mean body temperature (degrees Fahrenheit) upon waking and is found to be (97.63, 98.17). We are also told that the standard error is 0.125.

a) Calculate an estimate of the population mean body temperature upon waking.

Estimate is midpoint of confidence interval:

```
In [1]: (97.63 + 98.17)/2
```

```
Out[1]: 97.9
```

b) Show python commands to calculate the critical value (t_q) for this interval, including any imports.

```
from scipy.stats import t
t.ppf(0.975, df=15)
```

c) Based on the information given, calculate the sample standard deviation of body temperature in the data.

$$se = \frac{s}{\sqrt{16}} = 0.125 \quad \Longrightarrow \quad s = (0.125)(4) = 0.50$$

For the following questions, we test whether these data are consistent with the typical assumption that normal human body temperature is 98.6 degrees Fahrenheit against a two-sided alternative hypothesis.

d) State the null and alternative hypotheses, defining any parameters used in the statements.

Define μ to be the population mean body temperature upon waking.

$$H_0 : \mu = 98.6 \quad \text{against} \quad H_A : \mu \neq 98.6$$

e) Calculate the value of the test statistic.

$$t = \frac{97.9 - 98.6}{0.125} = -5.6$$

```
In [2]: (97.9 - 98.6)/0.125

Out[2]: -5.599999999999909
```

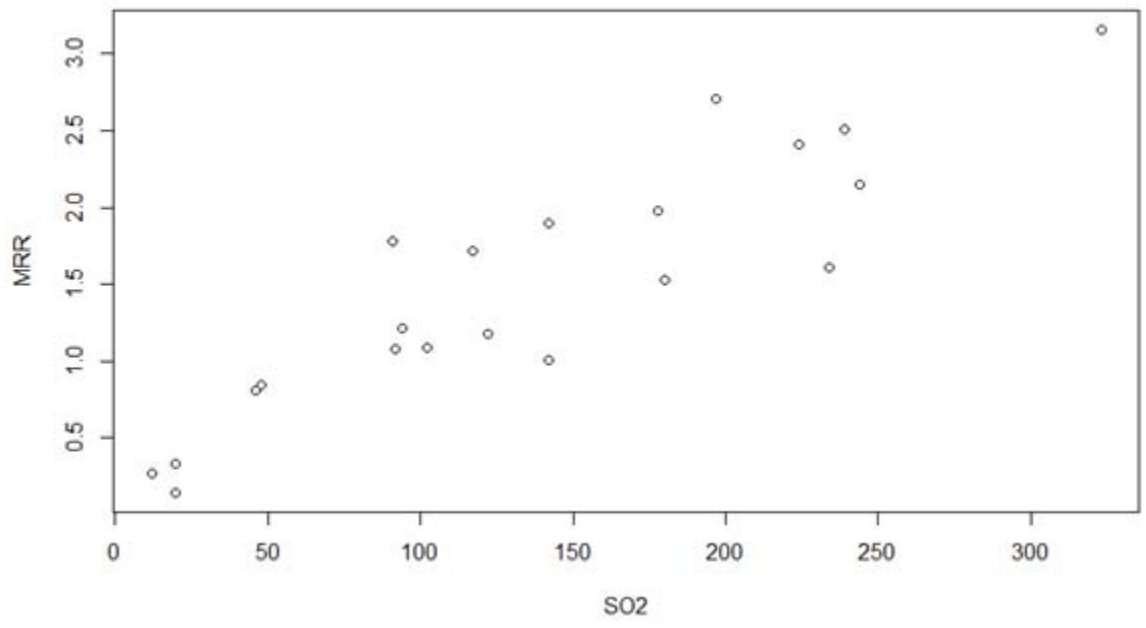
f) Based on the available information, do you reject or accept the null hypothesis at a signficance level of 0.05? Explain briefly.

The 95% confidence interval does not contain the null value 98.6, therefore we reject H_0 at the 0.05 level.

Notice that you could deduce the value of the 0.05 cutoff for t from the available information: divide the halfwidth of the confidence interval by the standard error. The result is the 0.05 cutoff for the absolute value of the t-test statistic.

---

2. Here is a scatter plot of Marble Tombstone Mean Surface Recession Rates (MRR) and Mean SO2 concentrations (SO2) over a 100-year period. (Source: T. C. Meierding (1993)).



The model summary for the linear regression of MMR on SO2 includes the following information:

	coef	std err	t	$P > \|t\|$	$[0.025$	$0.975]$
Intercept	0.3229959	0.1521958	2.122	0.0472	**	**
SO2	0.0085933	0.0009499	9.046	2.58e-08	**	**

a) Based on the graph, select the closest value for the sample correlation coefficient between MRR and SO2.

```
i. 1.2      ii. 0.9      iii. 0      iv. -0.5
```

The answer is ii. We can eliminate i. (correlations cannot be > 1), iv. (slope is not negative), and iii. (there is a pretty strong positive linear assocation).

It is also informative to realize here that the sample correlation between SO2 and MMR for this model is the square root of R^2, the proportion of variance explained.

b) Using the estimates above, write down the formula for the final regression model (least square line formula).

Let y denote body temperature. Round to 3 significant digits. Fitted model is

$$\hat{y} = 0.323 + 0.00859 * SO2$$

c) If I measure an SO2 concentration of 2 units, what recession rate would you predict based on the model?

$$\hat{y} = 0.323 + 0.00859 * 2 = 0.340$$

```
In [3]: 0.323 + 0.00859 * 2
Out[3]: 0.34018000000000004
```

d) For each 1 unit increase of SO2 concentration in the marble, how much increase/decrease of recession rate would you expect on average?

0.00859

e) If we use a critical value t_q = 2 for a 95% confidence interval, compute the confidence interval for the slope coefficient.

margin of error = (2)(0.000950) = 0.0019

confidence interval: (0.00859 - 0.0019, 0.00859 + 0.0019) = (0.00669, 0.01049)

f) I want to perform a hypothesis test on the slope to see if it is zero or not. What would be the value of the test statistic and the p-value in this case?

test statistic value is t = 9.046

p-value is 2.58 x 10**(-8)

g) Would you say there is statistical evidence that SO2 concentration and surface recession rates are indeed linearly associated, at a significance level of 0.05?

Yes, because the p-value is much smaller than 0.05.

3. We ran a regression for a dataset that studied human happiness across different nations. The response was LSI, which is the life-satisfaction index. The explanatory variables include GINI, CORRUPT, and DEMOCRACY. GINI is the measure of the inequality in income distributions (higher number means greater inequality). CORRUPT is the degree of corruption in the government (higher number means less corruption). DEMOCRACY is the degree of democracy (higher number means more civil and political liberties).

OLS Regression Results

Dep. Variable:	LSI	R-squared:	0.555
Model:	OLS	Adj. R-squared:	0.471
Method:	Least Squares	F-statistic:	6.649
Date:	Wed, 06 Nov 2019	Prob (F-statistic):	0.003997
Time:	10:42:26	Log-Likelihood:	**
No. Observations:	20	AIC:	**
Df Residuals:	16	BIC:	**
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.3404	1.5726	2.124	0.0496	**	**
GINI	0.0210	0.0336	0.625	0.5406	**	**
CORRUPT	0.2672	0.1330	2.010	0.0616	**	**
DEMOCRACY	0.1852	0.2138	0.866	0.3992	**	**

a) Which of the explanatory variables are significant at a 5% significance level?

None of the variables is significant at the 0.05 level based on their coefficient p-values.

CORRUPT is closest, with p=0.0616.

b) Show how to calculate the predicted LSI value for a nation with GINI = 42.7, CORRUPT = 22, and DEMOCRACY = 36.2. Just set it up, you don't need to do the calculation.

$$\hat{y} = 3.3404 + 0.0210 * 42.7 + 0.2672 * 22 + 0.1852 * 36.2$$

c) What is the proportion of variance in the response explained by these explanatory variables?

This is given by R-squared = 0.555

d) A small P-value for the F-statistic reported for the model implies that all explanatory variables are significantly different from 0. Circle only one answer.

TRUE FALSE

False, it implies that at least one of their coefficients is not zero.

e) In order to compute a 95% confidence interval for the coefficient of CORRUPT, we need to find the multiplier for the standard error. Show how to use the scipy.stats.t to compute this multiplier, including importing necessary package(s) and inserting numerical values for the arguments.

We need t_q for a confidence level of 0.95. For degrees of freedom use 'DF residuals' from the summary.

```
from scipy.stats import t
t.ppf(0.975, df=16)
```

---

4. For each of the following questions, assume we have already run the following commands:

```
In [4]: import numpy as np
        from scipy.stats import norm, t
```

a) Suppose X is normally distributed with population mean=5 and standard deviation=2. What Python commands are needed to compute the probability that X is larger than 8?

```
1 - norm.cdf(8, loc=5, scale=2)
```

b) With X the same as in a), what Python commands are needed in order to compute the probability that $|X-5| > 3$?

```
2*norm.cdf(-3, loc=0, scale=2)
```

c) Will the output of t.ppf(0.73, df = 56) be larger or smaller than 0?

Larger than 0 because the 73rd percentile is larger than the median, which equals 0

d) Will the output of norm.ppf(0.73, loc = -5, scale = 0.001) be larger or smaller than 0?

Less than zero because the small scale ensures that the 73rd percentile is still close to -5.

It is less than the 95th percentile, which is less than -5 + (2)(0.001) = -4.998

e) Which of the following commands will give the largest value:

```
norm.cdf(-1, loc=0, scale=1)
t.cdf(-1, df=3)
t.cdf(-1, df=30)
```

All three distributions have the same center at 0. The t with 3 df has the heaviest tails so the answer is

```
t.cdf(-1, df=3)
```

f) Which of the following will give you the 45th percentile of the standard normal distribution?

```
norm.cdf(45)
norm.cdf(0.45)
norm.ppf(0.45)
norm.ppf(0.475)
```

norm.ppf(0.45)

---

5. A sample survey interviews a simple random sample of 300 college women. It is hypothesized that 60% of all college women have been on a diet during the past year.

a) If $\hat{p}$ is the proportion of women in the sample that have been on a diet in the past year, find the numerical values for the mean and the standard deviation of the sampling distribution for $\hat{p}$, assuming the hypothesis is true.

$$E(\hat{p}) = 0.60 \qquad sd(\hat{p}) = \sqrt{\frac{(0.6)(0.4)}{300}} = 0.028$$

```
In [5]: import numpy as np
        np.sqrt(0.6*0.4/300)
```

```
Out[5]: 0.0282842712474619
```

b) If in the sample $\hat{p} = 0.54$, compute the z-test value for testing the hypothesis against a two-sided alternative. Does the value exceed 1.96 in absolute value?

We can use the null hypothesis value for $sd(\hat{p})$, and compute

$$z = \frac{0.54 - 0.60}{0.028} = -2.14$$

Yes, |-2.14| > 1.96.

```
In [6]: (0.54-0.60)/0.028
```

```
Out[6]: -2.1428571428571406
```

c) Calculate the margin of error for 95% confidence interval for the population proportion p. Use $z_q = 1.96$.

First calculate the standard error:

$$se = \sqrt{\frac{(0.54)(0.46)}{300}} = 0.029$$

The margin of error is (1.96)(0.029) = 0.056

In [7]:
```
se = np.sqrt(0.54*0.46/300)
me = 1.96*se
(se, me)
```

Out[7]:
```
(0.02877498913987632, 0.05639897871415758)
```

---

6. A gene expression diagnostic test for breast cancer yields the following results for 100 women with known cancer status.

	Cancer	No Cancer
Test positive	34	8
Test negative	6	52

a) The sensitivity of a test is the probability of testing positive if the person has cancer. Using the data, calculate an estimate of the sensitivity of this test.

In [8]:
```
34 / (34 + 6)
```

Out[8]:
```
0.85
```

b) The specificity of a test is the probability of testing negative if the person does not have cancer. Using the data, calculate an estimate of the specificity of this test.

In [9]:
```
52 / (8 + 52)
```

Out[9]:
```
0.8666666666666667
```

c) Calculate the odds-ratio for cancer versus no cancer for those who test positive versus those who test negative. Does the value support the idea that the test is more effective than random guessing, at least?

In [10]:
```
(34/8)/(6/52)
```

Out[10]:
```
36.83333333333333
```

Random guessing => odds ratio =1. The value here is much higher than 1 so much better than random guessing.

d) The standard error for the sample log-odds-ratio is

$$\sqrt{\frac{1}{34}+\frac{1}{8}+\frac{1}{6}+\frac{1}{52}}=0.583$$

Calculate the z test statistic if we wish to test the null hypothesis that the populatoin log-odds-ratio is zero.

In [11]:
```
z = np.log(36.833)/0.583
z
```

Out[11]:
```
6.185924841295014
```

---

7. Consider the following code and results for data from the Galapagos islands.

In [12]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns; sns.set()
import statsmodels.api as sm
import statsmodels.formula.api as smf
import statsmodels.regression.linear_model as lm
```

In [13]:
```
gala = pd.read_csv("gala.csv")
gala.head(3)
```

Out[13]:

	Island	Species	Endemics	Area	Elevation	Nearest	Scruz	Adjacent
0	Baltra	58	23	25.09	346	0.6	0.6	1.84
1	Bartolome	31	21	1.24	109	0.6	26.3	572.33
2	Caldwell	3	3	0.21	114	2.8	58.7	0.78

In [14]:
```
model1 = smf.ols('Species ~ Area + Elevation + Nearest + Adjacent', data=gala)
.fit()
model0 = smf.ols('Species ~ Area + Elevation', data=gala).fit()
f, p, df = model1.compare_f_test(model0)
pd.DataFrame({'f': [f], 'pvalue': [p], 'df_diff': [df]})
```

Out[14]:

	f	pvalue	df_diff
0	10.131299	0.000599	2.0

a) Write down the mathematical form of the regression model for 'model1' using generic symbols for the regression coefficients and defining the explanatory and response variables.

$$E(Species) = \beta_0 + \beta_1 * Area + \beta_2 * Elevation + \beta_3 * Nearest + \beta_4 * Adjacent$$

b) Using the same mathematical notation, state the null hypothesis being tested by the F test reported in the output of 'model1.compare_f_test(model0)'.

The null hypothesis is that model2 is adequate, that is:

$$H_0 : \beta_3 = \beta_4 = 0$$

c) based on the F test result which model do you choose and why?

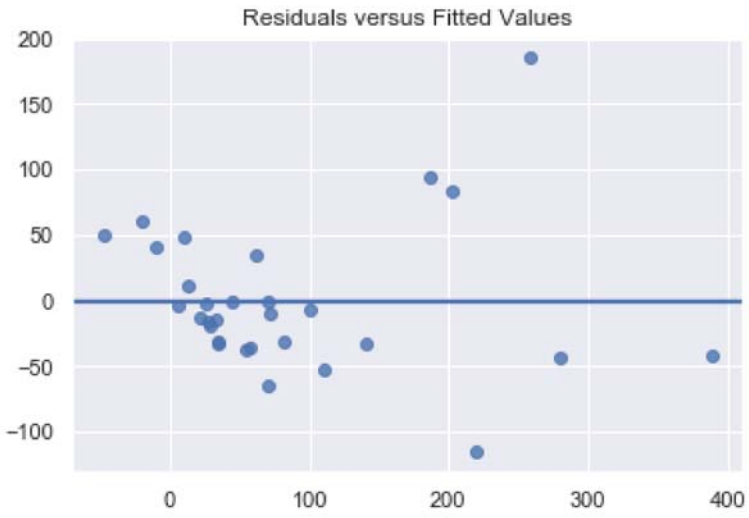
The p-value for the test is less than 0.001, so I reject H_0 and choose 'model1'.

d) Provide the additional python code needed if we want to see the coefficient estiamtes and their standard errors for 'model1'.

```
model1.summary()
```

e) Below is the residual plot for 'model1'. Recall that the ordinary least squares (OLS) fitting method is efficient if the model errors have constant standard deviation and the model is correct, with no nonlinear trends. Do the residuals appear consistent with these assumptions? Why or why not?

In [15]:
```
sns.regplot(x=model1.fittedvalues, y=model1.resid, ci=None)
plt.title('Residuals versus Fitted Values')
plt.show()
```



There are a couple of problems:

- The spread appears to increase as the fitted values increase; nonconstant error standard deviation;
- There seems to be some systematic curvature to the shape of the residual cloud as a function of the fitted values; suggests a missing variable or need for soem kind of transformation of the response.

8. Consider the following python code fragments and output from some analysis.

```
dat.head()

Out:
          x1       x2  y
    0   0.343687    -1  0
    1   1.848400    -1  1
    2   0.224359    -1  0
    3   -1.633660   -1  0
    4   1.245538    -1  1
```

```
dat['x2'].value_counts()

Out:

     1    50
    -1    50
    Name: x2, dtype: int64
```

```
mod = smf.logit('y ~ x1 + x2', data=dat).fit()
print(mod.summary())

Out:

Optimization terminated successfully.
Current function value: 0.447432
Iterations 7
                          Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  100
Model:                          Logit   Df Residuals:                       97
Method:                           MLE   Df Model:                            2
Date:                Thu, 07 Nov 2019   Pseudo R-squ.:                  0.3498
Time:                        12:31:49   Log-Likelihood:                -44.743
converged:                       True   LL-Null:                       -68.814
Covariance Type:            nonrobust   LLR p-value:                 3.517e-11
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.4948      0.276     -1.794      0.073      -1.035       0.046
x1             1.8939      0.382      4.964      0.000       1.146       2.642
x2             0.3683      0.274      1.345      0.179      -0.168       0.905
==============================================================================
```

a) How many observations (rows) are there in the data frame?

n=100 based on either the .value_counts() total or the reported "No. Observations."

b) Which of the logistic regression coefficients are significantly different from zero, at the significance level $\alpha = 0.05$?

Only the coefficient for x1, with p-value < 0.05.

c) Based on the fitted model, calculate an estimate of $p = P(Y = 1)$ if $X_1 = 0$ and $X_2 = 1$.

- First we estimate

$$log(odds) = -0.4948 + 1.8939 * 0 + 0.3683 * 1 = 0.3683 - 0.4948 = -0.1265$$

- Then we calculate $odds = e^{-0.1265} = 0.8812$.
- Finally we calculate

$$p = \frac{0.8812}{1 + 0.8812} = 0.468$$

d) Calculate the estimated log-odds-ratio for y=1 versus y=0 for individual A versus Individual B with the following values for x1 and x2:

Individual	x1	x2
A	2.0	1
B	2.0	-1

This is the difference between the fitted log-odds for the two individuals. The intercept and x1 terms cancel so we are left with:

$$(1)(\hat{\beta}_2) - (-1)(\hat{\beta}_2) = 2 * 0.3683 = 0.7366$$

e) Give a 95% confidence interval for the odds ratio for y=1 versus y=0 for individual A versus individual B above.

In [16]:
```
## We need to exponentiate 2 times the given confidence interval endpoints for
$\beta_2$:

(np.exp(-0.168*2),  np.exp(0.905*2))
```

Out[16]:
```
(0.7146231058160573, 6.11044743223061)
```