**Quantitative Political Methods**
**Final Exam Review**
*****This list does not include terms and concepts from the first half of the semester—remember that the final will include some questions from material we covered before the last exam!**

## Terms

**Association and Causality**

- Association
- Time order
- Direct causal relationship
- Spurious relationship
- Chain relationship
- Multiple causation
- Direct and indirect causation
- Treatment/control variables
- Explanatory/response variables
- Counterfactual
- Average Treatment Effect

**Comparing Means and Proportions**

- Dichotomous/binary variable
- Conditional probabilities
- Conditional distribution
- Pooled variance/standard deviation
- Dependent samples/matched pairs
- Statistical dependence/independence
- Contingency table
- Observed/expected frequency
- Chi squared distribution

**Regression**

- Linear regression
- Regression analysis
- Y-intercept
- Slope
- Least squares line
- Sum of squared error
- Conditional distribution
- Conditional standard deviation
- Unconditional standard deviation
- Mean squared error/root mean squared error
- Total sum of squares
- Correlation coefficient (r/R)
- Coefficient of determination ($r^2$)
- Statistical control

**Regression, cont.**

- Control variable
- Statistical interaction
- Confounding variables
- Multiple regression function
- Simpson's paradox
- Multiple correlation ($R^2$)
- F-statistic
- Interaction
- Dummy variable
- Partial sums of squares
- Autocorrelation
- Multicollinearity
- Heteroscedasticity
- Leverage
- Outliers

**Causality**

- Average treatment effect
- Counterfactual
- Causality
- Experiment
- Difference-in-differences
- Regression discontinuity
- Instrument
- Instrumental variables

## Concepts

### Comparing Means and Proportions

- Construct a confidence interval for difference in proportions
- Construct a confidence interval for difference in means
- Conduct a hypothesis test for difference in proportions
- Conduct a hypothesis test for difference in means (large sample)
- Conduct a hypothesis test for difference in means (small sample)
- Calculate pooled standard error for means (large sample)
- Calculate pooled standard error for means (small sample)
- Calculate a pooled estimate for two proportions
- Calculate a pooled standard error for two proportions
- Interpret p-values for differences in means and proportions
- Calculate expected frequencies for a contingency table
- Calculate a Chi square statistic
- Conduct a hypothesis test and interpret the p-value for a Chi square statistic
- Calculate standardized residuals for a contingency table

### Regression

- Interpret y-intercept and slope for a linear function
- Calculate $\alpha$ and $\beta$ and write a prediction equation
- Calculate Sum of Squared Errors
- Calculate Mean Squared Error/ Root Mean Squared Error (conditional variance/standard deviation)
- Calculate Total Sum of Squares
- Interpret a scatter plot
- Construct a confidence interval around $\beta$
- Conduct a hypothesis test for $\beta$
- Calculate standard error for $\beta$
- Calculate $S_x$ and $S_y$
- Calculate r and interpret its meaning
- Calculate $r^2/R^2$ for linear and multiple regressions and interpret its meaning
- Calculate unconditional standard deviation
- Interpret a regression analysis table
- Write a prediction equation for a multiple regression function
- Interpret regression coefficients in a multiple regression
- Interpret the significance of a multiple regression using the F distribution
- Test an interaction term
- Interpret a regression line with an interaction term
- Plot regression lines

- Interpret error terms in regression equations
- Interpret residuals plotted together

**Causality**

- Why do experiments help us estimate causal effects
- Assumptions needed to make causal inference using regression in observational (non-experimental) data.
- Interpret the results from a difference-in-differences model
- Interpret the results from a regression discontinuity regression
- Interpret results of instrumental variables regression
- Understand the assumptions important for making causal inference in diff-in-diff, RD, and IV models

**Practice Problems**

1. A survey asked 683 students how many hours they spent doing homework in a week. On average, the 292 surveyed men responded 8.4 hours. The standard deviation of their responses was 9.5. On average, the 391 surveyed women responded 12.8 hours, with standard deviation of 11.6. We are interested in testing whether men spend less hours on homework than women. (Note: round all numbers to the nearest thousandth).

    a. Formulate the null and alternative hypotheses.

    b. Calculate the standard error for this hypothesis test

    c. With a confidence level of 99%, what would you conclude?

d. Construct a 99% confidence interval for the difference in the two means. Interpret.

e. What is the relationship between the p-value you observed in part c, and the confidence interval calculated in part d?

2. A group of researchers are conducting experiments to see how the price of mosquito nets sold at hospitals affects the number of patients who choose to purchase them. Unfortunately, the researchers have had computer trouble, and they have only been able to retrieve the information presented in the following table. It contains partial information for each cell, including some observed counts, some expected frequencies (in parentheses), and some column and row totals.

| | Price of nets | | | |
|---|---|---|---|---|
| | $0.65 | $1.60 | $2.50 | Total |
| Purchased net | 166 (121.27) | | 92 | |
| Did not purchase | 44 | 104 (84.50) | | 259 |
| Total | | 200 | | |

a. Use the information listed to complete the table. Be sure to calculate both the observed and expected frequencies for each cell.

b. Calculate the standardized residual for the upper-left cell of the table (i.e., Purchased nets when the price was $0.65). Interpret this statistic.

c. Calculate the $\chi^2$ statistic for this table. Specify and conduct a hypothesis test using this number. Interpret your results. What do these results tell us about the relationship between price and the purchasing of nets?

3. For the multiple regression output below, n = 50.

```
Call:
lm(formula = Y ~ X1 + X2 + X3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -272.028    253.244  -1.074    0.288
X1                      305.994   0.487
X2             3.216     16.369
X3            45.971      3.171  14.495   <2e-16

RMSE:
R-squared:
F-statistic:



 (TSS <- sum((Y - mean(Y))^2))
105,176,812
 (SSE <- sum(mod1$residuals^2))
18,198,652
```

Find:

a. The p-value for the effect of X1 on Y

b. The test statistic for the null hypothesis that X2 has no effect on Y. Find the p-value and interpret.

c. The RMSE

d. The $R^2$ value

e. A 95% confidence interval for the estimated effect of X3 on Y

f. The X1 estimate

g. The F Statistic. Calculate the p-value and interpret.

h. Write the multiple regression equation.

4. In 2007, 11 South American countries had an average GDP per capita of $10,752 with a standard deviation of $6,500. 14 Central American countries had an average GDP per capital of $12,936 with a standard deviation of $3,500. Are income levels different between these two regions?

5. The National Health Interview Survey estimated that current cigarette smokers were 41.9% of American adults in 1965 and 21.5% in 2003. In 1965, the sample size was 4,238 and in 2003 the sample size was 6,947.

    a. Does the percentage of smokers in 2003 differ from the percentage of smokers in 1965 at a significance level of 0.01?

b. Construct a 99% confidence interval around the difference in the two proportions.

6. Researchers interested in determining if there is a relationship between death anxiety and religiosity conducted the following study. Subjects completed a death anxiety scale (high score=high anxiety) and also completed a checklist designed to measure an individual's degree of religiosity (high score=greater religiosity. For simplicity's sake, assume that religiosity is the explanatory (independent) variable and death anxiety is the response (dependent) variable). Use the data provided to answer the questions below.

| X (religiosity) | Y (death anxiety) | $(X_i-X_{bar})$ | $(Y_i-Y_{bar})$ | $(X_i-X_{bar})^* (Y_i-Y_{bar})$ | $(X_i-X_{bar})^2$ | $(Y_i-Y_{bar})^2$ |
|---|---|---|---|---|---|---|
| 14 | 24 | 1.625 | 5.125 | 8.328 | 2.641 | 26.266 |
| 5 | 31 | -7.375 | 12.125 | -89.422 | 54.391 | 147.016 |
| 10 | 19 | -2.375 | .125 | -.297 | 5.641 | 0.016 |
| 18 | 6 | 5.625 | -12.875 | -72.422 | 31.641 | 165.766 |
| 19 | 8 | 6.625 | -10.875 | -72.047 | 43.891 | 118.266 |
| 4 | 38 | -8.375 | 19.125 | -160.172 | 70.141 | 365.766 |
| 14 | 14 | 1.625 | -4.875 | -7.922 | 2.641 | 23.766 |
| 15 | 11 | 2.625 | -7.875 | -20.672 | 6.891 | 62.016 |
| $X_{bar}$=12.375 | $Y_{bar}$=18.875 | | | **Total** -414.626 | **Total** 217.878 | **Total** 908.878 |

a. Calculate $\beta$.

b. Calculate $\alpha$.

c. Write the regression equation and substantively interpret each part of the equation.

d. Calculate conditional and unconditional standard deviation

e. Calculate TSS

f. Calculate SSE

g. Calculate $r$

h. Calculate $r^2$

i. Does a correlation exist between religiosity and death anxiety? Conduct a hypothesis test to find out.

7. Below is a logit function to model admission into UCLA graduate programs:

$$\text{Logit}[P(Y = 1)] = -4.949 + 0.0026*GRE + 0.757*GPA$$

Find the estimated probability of being admitted to graduate school for:

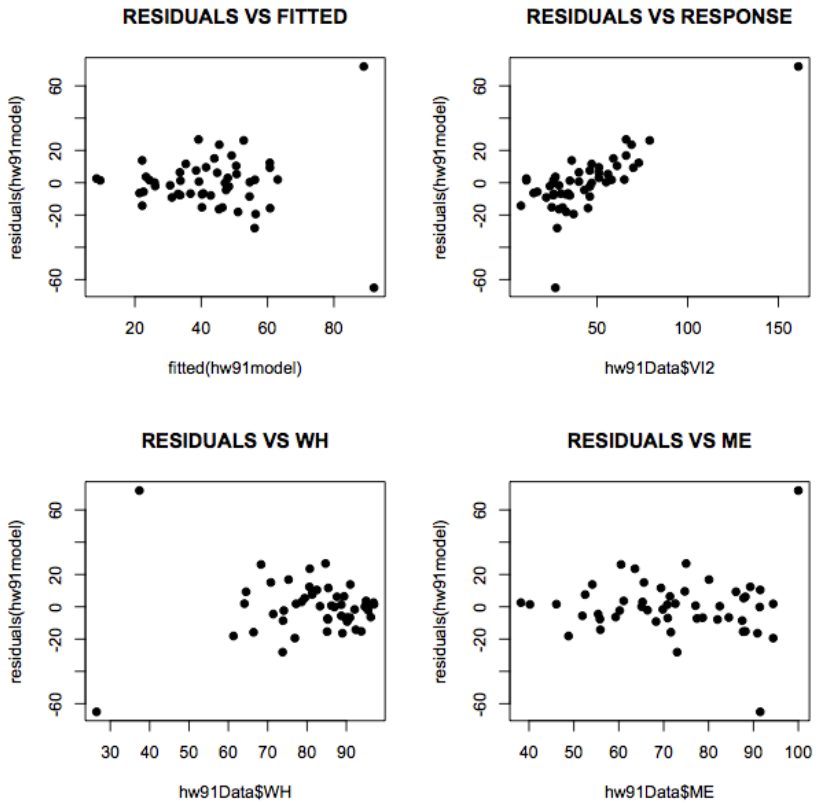
a. An applicant with a GRE score of 750 and a GPA of 4.0

b. An applicant with a GRE score of 450 and a GPA of 2.0

c. Conduct a hypothesis test for the GPA coefficient. Interpret the results.

8. In the plots below, the residuals of a regression model are plotted against several different quantities. The regression model predicts violent crime rates for states based on several covariates. WH is one of these covariates, as is ME (percent of people living in metropolitan areas).

   "Residuals vs Fitted" plots the residuals for the regression against the fitted (expected) values for each point. "Residuals vs Response" plots the residuals against the response (outcome) variable). "Residuals vs WH" and "Residuals vs ME" plot the residuals against each of the covariates.




   a) What is indicated by the positive relationship between the residuals and the response variable?

b) Is there evidence of any outliers? If so, do you think those outliers seem to be high leverage? Explain your answer.

c) Do these residuals indicate any other potential violations of standard regression assumptions?

9. A Pew Research Center poll (May 14, 2003) of 1201 adults asked, "All in all, do you think affirmative action programs designed to increase the number of black and minority students on college campuses are a good thing or a bad thing?" 60% said good, 30% said bad, and 10% didn't know. Does this data prove that a majority of people are in favor of affirmative action? Conduct a hypothesis test using a significance level of 0.05.

10. For recent data in Florida on y = selling price of a home (in dollars), $x_1$ = size of home (in square feet), $x_2$ = lot size (in square feet), the prediction equation is:

$$y = -10{,}536 + 53.8x_1 + 2.84x_2$$

a. A particular home of 1240 square feet on a lot of 18,000 square feet sold for $145,000. Find the predicted selling price and the residual, and interpret.

b. For fixed lot size, how much is the house's selling price predicted to increase for each square-foot increase in home size? Why?

11. The table below is a regression output for the selling price of 32 grandfather clocks sold at auction. "Age" represents age of the clock (in years), "Bidders" represents the number of individuals participating in bidding, and "Price" represents selling price (in pounds sterling).

| **Predictor** | **Coefficient** | **SE for Coefficient** |
|---|---|---|
| Constant | 322.8 | 293.3 |
| Age | 0.873 | 2.020 |
| Bidders | -93.41 | 29.71 |
| Age*Bidders | 1.2979 | 0.2110 |

a. Write the multiple regression equation.

b. Test if the "Age" coefficient is significantly different from 0 and interpret the results.

c. Construct a 92% confidence interval around the interaction coefficient and interpret.

d. Draw the regression line that would result in a hypothetical situation with 0 bidders. Hypothetically, what does the y-intercept on this graph represent?

e. If the age of a clock is 100 years, what happens as you increase the number of bidders?

f. If the age of a clock is 20 years, what happens as you increase the number of bidders? Compare this answer to part e—why does this happen? What does this mean about the regression equation?

12. Which of the following are measures of spread?

    a. Median
    b. Standard deviation
    c. Range
    d. Root mean squared error

13. X is a random variable with mean $\mu$ and variance $\sigma^2$. The "standardized form" of X is Z = (X- $\mu$)/ $\sigma$. What is the mean and variance, respectively, of Z?

    a. Mean = 0, variance = 1
    b. Mean = 1, variance = 0
    c. Mean = 0.05, variance = 0

14. A correlation coefficient of $r$ = 0.8 is reported for a sample of paris (x,y). Without any further information, this implies that…

    a. As the x values decrease, the y values increase.
    b. (x,y) are scattered about a straight line of unknown positive slope.
    c. 80% of the variation in y is due to regression on x.
    d. The points (x,y) are scattered about a straight line of slope 0.8.