

Practice Questions for Exam #2 (with solutions)

1. Suppose that we have collected a stratified random sample of 1,000 Hispanic adults and 1,000 non-Hispanic adults. These respondents are asked whether they would be willing to vote for New Mexico governor Bill Richardson for president in 2008. 721 of the Hispanic respondents responded that they would while 513 of the non-Hispanics respondents that they would. Hispanics make up 12.4% of the adult population.

(a) What is the sample proportion of respondents who would be willing to vote for Bill Richardson?

The sample proportion can be computed as $p = (721 + 513) / 2000 = 61.7\%$

(b) What is the weighted sample proportion of respondents who would be willing to vote for Bill Richardson?

The weighted sample proportion can be computed as,

$$p_w = \underbrace{0.124}_{\substack{\text{Fraction} \\ \text{of Hispanics} \\ \text{in the Population}}} * \underbrace{\frac{721}{1000}}_{\substack{\text{Fraction} \\ \text{of Hispanics} \\ \text{who would} \\ \text{vote for} \\ \text{Richardson}}} + \underbrace{(1-0.124)}_{\substack{\text{Fraction} \\ \text{of non-Hispanics} \\ \text{in the Population}}} * \underbrace{\frac{513}{1000}}_{\substack{\text{Fraction} \\ \text{of non-Hispanics} \\ \text{who would} \\ \text{vote for} \\ \text{Richardson}}} = 53.9\%$$

- (c) Which of these estimators provides an accurate estimate of the population proportion of American adults who would be willing to vote for Bill Richardson. *Only the weighted sample proportion produces an unbiased estimate of the population proportion of American adults who would be willing to vote for Richardson, because it takes into account the fact that Hispanics were more likely to be included in the sample.*

2. Suppose that our class was hired as consultants to conduct an exit poll for the last U.S. Senate race in New York. We polled voters at selected precincts in order to determine whether they voted for Hillary Clinton or John Spencer for Senate. We found that 61% of male voters voted for Clinton while 73% of Female voters voted for Clinton. We know in fact that 50% of the voting population in New York is Female, but we found that 60% of the sample is Female.

- (a) What is the sample proportion of New York voters who voted for Hillary?

The sample proportion can be computed as,

$$p = \underbrace{0.4}_{\substack{\text{Fraction} \\ \text{of Males} \\ \text{in the Sample}}} * \underbrace{0.61}_{\substack{\text{Fraction} \\ \text{of Males} \\ \text{who voted} \\ \text{for Clinton}}} + \underbrace{0.6}_{\substack{\text{Fraction} \\ \text{of Females} \\ \text{in the Sample}}} * \underbrace{0.73}_{\substack{\text{Fraction} \\ \text{of Females} \\ \text{who voted} \\ \text{for Clinton}}} = 68.2\%$$

- (b) Does the sample proportion provide an unbiased estimate of the population in this case? If so, what could be causing this bias? What could be done to correct for

this bias? Compute a corrected estimate for the population proportion of voters who choose Hillary.

It seems as if there are far too many females in our sample. This is likely due to higher response rates among female voters to the exit poll survey. We can correct for this problem by weighting by gender. We can compute the weighted sample proportion as,

$$p_w = \underbrace{0.5}_{\substack{\text{Fraction} \\ \text{of Males} \\ \text{in the Population}}} * \underbrace{0.61}_{\substack{\text{Fraction} \\ \text{of Males} \\ \text{who voted} \\ \text{for Clinton}}} + \underbrace{0.5}_{\substack{\text{Fraction} \\ \text{of Females} \\ \text{in the Population}}} * \underbrace{0.73}_{\substack{\text{Fraction} \\ \text{of Females} \\ \text{who voted} \\ \text{for Clinton}}} = 67.0\%$$

Since we had too many female voters in our sample and these female voters were more likely to vote for Hillary, we were overstating her support with the sample proportion.

3. The sample correlation between X and Y is $r_{XY} = 6\%$. This correlation is computed from a sample of $N = 251$ observations.

(a) Is the correlation statistically significant at the 5% level?

We consider the hypothesis test $H_0 : \rho = 0$. The Z-statistic for this null hypothesis

is, $Z = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = 0.9504$, hence the correlation is not statistically significant at

the 5% level.

(b) Is the correlation statistically significant at the 1% level?

The correlation coefficient is not statistically significant at the 1% level.

(c) Suppose that X has a mean of 2 and a standard deviation of 1 and Y has a mean of -3 and a standard deviation of 0.5. What are the intercept and slope coefficients

that would result from a bivariate regression with Y as the dependent variable and X as the independent variable.

Recall that the formulas for these coefficients are $b_1 = r \frac{s_y}{s_x} = 0.03$ and

$$b_0 = \bar{Y} - b_1 \bar{X} = -3.06$$

4. The following regression results were generated using the mba admissions dataset we considered in class.

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-64.456	11.307		-5.700	.000
	Quality of Admissions Essay	6.182	2.477	.091	2.496	.013
	Quality of Letters of Recommendation	8.471	1.967	.158	4.307	.000
	Female	2.525	1.686	.056	1.498	.135
	Quantitative GMAT	.998	.131	.284	7.637	.000
	Verbal GMAT	.748	.115	.240	6.501	.000
	Accounting, Business, or Economics major	1.963	2.669	.053	.735	.462
	Math, Science, or Engineering Major	4.924	2.577	.139	1.911	.057

a Dependent Variable: Percentile Rank at Graduation

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.443(a)	.196	.187	15.68421

- (a) Interpret the coefficients in this regression.

Holding the other independent variables constant, if the quality of the admissions essay increases by one unit, the student graduating rank is expected to rise by 6

percentage points. If the quality of the letters of recommendation increases by one unit, the student rank is expected to increase by 8 percentage points. Females are expected to graduate 2.5 percentage points higher in rank. If quantitative GMAT increases by 1, then rank is expected to increase by one percentage point. If verbal GMAT increases by 1, then rank is expected to increase by 0.7 percentage points. Accounting, Business, and Economics majors are expected to graduate 2 points higher than those that didn't major in Accounting, Business, Economics, Math, Science, or Engineering. Math, Science, and Engineering majors are expected to graduate 4.9 percentage points higher than non-Accounting, Business, Economics, Math, Science, Engineering majors.

- (b) Does undergraduate major have an effect on graduating rank? Use a 5% significance level.

Neither coefficient relating to major is statistically significant at the 5% level, hence, we don't find any evidence that undergraduate major has an effect on graduating rank.

- (b) Form a 95% confidence interval for the coefficient on 'gender'. Interpret this confidence interval.

*We can form a confidence interval for gender using $2.525 \pm 1.96 * 1.686 = [-0.780, 5.830]$. Since 0 is contained in this interval, we fail to reject the null hypothesis $H_0 : \beta_{Gender} = 0$. Hence, we determine that gender does not have a statistically significant effect.*

- (c) Suppose we are interested in testing whether Quantitative GMAT score has a significant effect on graduating percentile rank. What would the appropriate null and alternative hypothesis be?

The appropriate null hypothesis would be $H_0 : \beta_{Quantitative} = 0$ and the appropriate alternative would be $H_A : \beta_{Quantitative} \neq 0$

5. Suppose that we are interested in determining what factors increase the likelihood that an individual will contribute money to a charitable organization. The following regression was estimated using data from the 1996 General Social Survey. The dependent variable, MoneyContributed, is equal to one if the individual contributed money to a charitable organization and zero otherwise. In the following regression, a linear probability model is used to predict the probability that individuals give to charity.

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.198	.099		-2.001	.046
	Age	.003	.004	.095	.793	.428
	Age^2	-1.25E-005	.000	-.042	-.356	.722
	Female	.018	.019	.019	.945	.345
	Black	-.040	.030	-.029	-1.332	.183
	Number of Children	.001	.007	.004	.151	.880
	Total Family Income (Categories)	.022	.004	.111	5.024	.000
	South	-.006	.022	-.005	-.250	.803
	Democrat	.002	.023	.002	.070	.944
	Republican	.037	.025	.035	1.492	.136
	Left-Right Ideology (1-7 Scale)	.012	.007	.035	1.621	.105

Number of Years of Schooling Completed	.008	.004	.048	2.178	.029
---	------	------	------	-------	------

Dependent Variable: MoneyContributed

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.166(a)	.027	.023	.46552

- (a) Interpret the coefficients in this regression.

The effect of age on the probability of charitable giving decreases as age decreases (though the effect is not statistically significant). Female are predicted to be 1.8% more likely to donate, holding the other independent variables constant. Blacks are 4.0% less likely to contribute. Each child in the household increases the probability of charitable giving by 0.1%. An increase in one income category increases the probability of charitable giving by 2.2%. Southerners are 0.6% less likely to contribute. Democratic party identifiers are 0.2% more likely to contribute than respondents who don't identify with either party. Republicans are 3.7% more likely to contribute than respondents who don't identify with either party. Moving one unit to the right on the self-placement scale increases the probability of charitable giving by 1.2%. Each year of schooling completed increases the probability of charitable giving by 0.8%.

- (b) Which coefficients are statistically significant at the 5% level? What about the 10% level?

Family income and number of school years completed are significant at the 5% level. These variables as well as Black are statistically significant at the 10% level.

- (c) Suppose our null hypothesis is that ideology does not affect the likelihood of charitable giving. How would we set up the null and alternative hypotheses?

The null hypothesis is given by, $H_0 : \beta_{Ideology} = 0$

and the alternative hypothesis is, $H_A : \beta_{Ideology} \neq 0$

- (d) Do you agree or disagree with the following statement: “holding everything else constant, increasing age by one year increases the predicted probability of charitable giving by 0.3%”? Explain.

This statement is not accurate- it is impossible to hold Age² constant while varying age. We must interpret the effect of age by considering both the Age and Age² terms. Here, we have a non-linear relationship where the effect of age on the dependent variable varies with the value of age itself. What we do know (from the fact that the Age² coefficient is negative) is that as we increase age, the effect of age on the dependent variable decreases.

6. In the following regression model, use the data file ‘wages_full_time.sav’. Run a regression with log of imputed wage as the dependent variable and school years, male, age, and age squared as independent variables.

- (a) Interpret all the coefficients in this regression.

The regressions results are as follows:

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	.727	.177		4.106	.000
	School Years	.057	.006	.295	8.927	.000
	Male	.227	.040	.173	5.695	.000
	Age	.045	.008	.981	5.856	.000
	Age Squared	.000	.000	-.785	-4.691	.000

a Dependent Variable: Log of Imputed Wage

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.381(a)	.145	.142	.53699

a Predictors: (Constant), Age Squared, Male, School Years, Age

Holding the other independent variables constant, each year of schooling leads to a 5.7% increase in wages. Male workers are expected to earn 22.7% more than females. The quadratic term is zero to three decimal places, but we can see, based on the sign of the t-statistic, that the coefficient is negative. Hence, the effect of age on log(Wages) has an inverted U-shape.

- (b) Which coefficients are statistically significant at the 5% level? What about the 1% level?

All the coefficients are statistically significant at both the 5% and 1% levels.

- (c) What is the predicted log(wage) of a female worker, age 40, with 14 years of schooling?

We can predict the log of wage for this individual using,

$$\text{LogWage} = 0.727 + 0.057 * 14 + 0.227 * 0 + 0.045 * 40 + 0.000 * 40^2 = 3.325$$

Using SPSS, we find that $\text{LogWage} = 2.648$. These numbers differ because of rounding errors in the first calculation (which are quite large in this case).

- (d) Interpret the R-squared of this regression.

We would say that about 33% of the variation in LogWages is explained by the variation in the independent variables. Consequently, the independent variables are reasonably useful in predicting wages.

- (e) Would you feel comfortable interpreting the coefficient on Gender as the causal effect of gender on wages? Explain.

In order to interpret the coefficient on gender as a causal effect, we would want to know that we are controlling for all important difference between the males and females in the sample relating to their value in the work place. In this case, we are not controlling for experience. Male and Female workers are expected to differ in experience, because more females than males will take time off from work in order to raise children, and later return to the workforce. We would therefore expect the Male coefficient to overestimate the causal effect of gender on wages.

- (f) Now, run the same regression, including also an interaction of male and school years. Interpret the coefficient on this interaction term.

Including the interaction term, we find,

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	.627	.230		2.731	.006
	School Years	.064	.012	.333	5.187	.000
	Male	.338	.167	.257	2.023	.043
	Age	.046	.008	.996	5.893	.000

Age Squared	.000	.000	-.800	-4.739	.000
male_educ	-.009	.014	-.093	-.682	.495

a Dependent Variable: Log of Imputed Wage

The interaction term tells us that the effect of education on wage is less for males than it is for females. For females, one extra year of education increases log of wages by 0.064 units (or increases wages by 6.4%). For males, one extra year of education increases log of wages by $0.064 - 0.009 = 0.055$ unit (or increases wages by 5.5%).

7. Suppose that we would like to determine whether democracies in the Northern hemisphere have higher turnout rates than democracies in the Southern hemisphere. Rather than using an independent samples test, we would like to use linear regression.

(a) What should we use as our dependent and independent variables?

We would create a dummy variable Northern, that is equal to 1 for all democracies in the Northern hemisphere and equal to zero for all democratic in the Southern hemisphere (alternatively, we could create a dummy variable Southern). This dummy variable would be our independent variable and our dependent variable would be turnout. Hence, our model would be

$$\text{Turnout}_n = \beta_0 + \beta_1 \text{Northern}_n + \varepsilon_n$$

(b) State the null and alternative hypothesis in terms of the mean turnout rate in Northern hemisphere and Southern hemisphere democracies.

If μ_N is the population mean turnout rate for Northern democracies and μ_S is the population mean turnout rate for Southern democracies, we would state the null and alternative hypotheses as $H_0 : \mu_N = \mu_S$ and $H_A : \mu_N \neq \mu_S$.

- (c) State the null and alternative hypothesis in terms of the regression coefficients.

The null and alternative hypotheses are $H_0 : \beta_1 = 0$ and $H_A : \beta_1 \neq 0$

- (d) We now suspect that the relationship between hemisphere and turnout depends on whether the country has compulsory voting. Specify a model of turnout that depends on hemisphere, compulsory voting, and an interaction term.

We have,

$$\text{Turnout}_n = \beta_0 + \beta_1 \text{Northern}_n + \beta_2 \text{Compusory}_n + \beta_3 \text{Northern}_n * \text{Compusory}_n + \varepsilon_n$$

- (e) What are the predicted turnout rates for each of the following 4 categories of countries: Northern hemisphere democracies without compulsory voting, Northern hemisphere democracies with compulsory voting, Southern hemisphere democracies without compulsory voting, and Southern hemisphere democracies with compulsory voting.

We have,

$$\mu_{NN} = \beta_0 + \beta_1 * 1 + \beta_2 * 0 + \beta_3 * 1 * 0 = \beta_0 + \beta_1$$

$$\mu_{NC} = \beta_0 + \beta_1 * 1 + \beta_2 * 1 + \beta_3 * 1 * 1 = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

$$\mu_{SN} = \beta_0 + \beta_1 * 0 + \beta_2 * 0 + \beta_3 * 0 * 0 = \beta_0$$

$$\mu_{SC} = \beta_0 + \beta_1 * 0 + \beta_2 * 1 + \beta_3 * 0 * 1 = \beta_0 + \beta_2$$

- (f) How would we write the null hypothesis that hemisphere matters in those countries without compulsory voting?

The null hypothesis would be $H_0 : \mu_{NN} = \mu_{SN}$, which can be written as

$$H_0 : \beta_0 + \beta_1 = \beta_0, \text{ or } H_0 : \beta_1 = 0$$

- (g) How would we write the null hypothesis that hemisphere does not matter in countries with compulsory voting?

We could write the null hypothesis as $H_0 : \mu_{NC} = \mu_{SC}$, which can be written as,

$$H_0 : \beta_0 + \beta_1 + \beta_2 + \beta_3 = \beta_0 + \beta_2, \text{ or equivalently, } H_0 : \beta_1 + \beta_3 = 0.$$

8. Suppose that we are interested in predicting the probability that an incumbent president wins re-election, two years before the election is to take place. Which of the following should we include as regressors?

- (a) Inflation in the first two years of the presidency.
- (b) The favorability rating of the challenger.
- (c) The number of hurricanes during the first two years of the presidency.
- (d) The presidents approval rating 6 months before the election.
- (e) GDP growth during the first two years of the presidency.

Variables (a) and (e) should be included as regressors because (i) they would be known two years before the election (when we are making the prediction) and are expected to be related to the likelihood that the incumbent is re-elected. Variable (b) and (d) would not be known when the prediction is being made, and variable (c) is unlikely to be related to the likelihood of re-election.

9. Suppose that we are interested in determining the effect of watching the Democratic convention on support for the Democratic presidential candidate. We

consider the following design. We randomly select 1,000 individuals to survey and ask them (i) whether they watched the debate and (ii) whether they support the Democratic candidate for president. We then compare the support for the president between the group that watched the debate and the group that did not. Is this a valid way to determine the causal effect of watching the convention coverage on support for the Democratic candidate? If yes, explain why. If not, explain why not and discuss how we could determine the causal effect.

This procedure will not yield the causal effect because individuals are choosing whether or not to watch the debate (self-selection). It is likely that the convention watching group differs from the non-watching groups in ways other than whether they watched the convention- e.g. the watchers are more likely to support democratic candidate a priori, they are more educated and interested in politics, they do not work at night, etc. Ideally, we could determine the causal effect by randomly assigning individuals to 'watch' and 'don't watch' groups. If this is not possible, then we would at minimum want to control for observed differences between these groups using multiple regression. For example, we could control for age, gender, education, and income. In addition, we could survey these individuals before and after the convention in order to determine whether there was any change in opinion in each group.

10. A Fox News poll of 900 likely voters conducted on 11/04-11/05 determined that the net approval rating (approve – disapprove) for President Bush was -16%. A CNN poll of 1008 American adults conducted 11/03-11/4 determined that the net

approval for President Bush was -26%. What could explain the difference between these two polls? Which poll is more accurate?

There are many reasons why these two polls differ. First there is sampling error. In addition, there are various sources of survey bias- unit non-response, item non-response, and measurement error. Though these surveys do not report response rates, most public opinion polls have very low response rates, and it is therefore likely that the respondents to each survey differ significantly from the population. Different survey organizations will word their presidential approval question differently leading to measurement error. Finally, these surveys have different target populations- likely voters vs. American adults.

It is difficult to say which survey is more accurate. CNN is a relatively new polling organization (until this year CNN polled with Gallup and USA Today) while Fox news has a more established track record. However, relying on American adults rather than likely voters is a more standard choice for measuring Presidential approval.

11. Our class has been hired as consultants to investigate discrimination at a Nissan car dealership. A group of young buyers claims that they have been offered systematically higher prices than older buyers. We use ‘invoice price’ to denote the price that the dealer pays for a car and let ‘sale price’ denote the price that a buyer pays. Typically, we would expect the dealer to charge a ‘mark-up’ of $x\%$ of the car price. Thus, if I_n denotes the invoice price of the car, $P_n = (1 + \alpha_n)I_n$ would denote the price the buyer pays, where α would be the markup. We can

rewrite this as $\log(P_n / I_n) = \log(1 + \alpha_n)$. Now let A_n denote the age of Buyer n .

In order to test for discrimination, we could let $\log(1 + \alpha_n) = \beta_0 + \beta_1 A_n + \varepsilon_n$

yielding the regression equation, $\log(P_n / I_n) = \beta_0 + \beta_1 A_n + \varepsilon_n$. Suppose that our goal is to determine whether there is age discrimination in markups. State the appropriate null hypothesis in terms of the coefficients of the model.

If $\beta_1 = 0$, this would imply that $\log(P_n / I_n) = \log(1 + \alpha_n) = \beta_0 + \varepsilon_n$, indicating that markups do not depend on age. Hence, the null hypothesis would be $H_0 : \beta_1 = 0$.

12. Suppose that I run a regression in order to predict voter turnout in a precinct based on some control variables. As control variables, I include R_n^1 (rain measured in inches) and R_n^2 (rain measured in millimeters). Is the ordinary least squares estimator properly defined in this case? If not, explain.

The ordinary least squares estimator will not be properly defined because of multi-collinearity. Essentially, R_n^2 is redundant if we have already included R_n^1 .

SPSS will deal with this problem by excluding one of the variables from the specification.