

Question 1: Many large corporations and government agencies administer a preemployment test in an attempt to screen job applicants. The test is supposed to measure an applicant's aptitude for the job and the results are used as part of the information for making a hiring decision. Data were collected on twenty job applicants, each of whom were hired on a trial basis for six weeks. One week was spent in a training class. The remaining five weeks were spent on the job. The participants were selected from a pool of applicants by a method that was not related to the preemployment test scores. A test was given at the end of the training period and a work performance evaluation was developed at the end of the six-week period. These two scores were combined to form an index of job performance, denoted y_i . Let X_i be the score on the preemployment test. Applicants were also classified into two racial groups, $Z_i = 1$ for minority applicants, and $Z_i = 0$ for white applicants. Regression analysis yielded the following results:

	Model 1		Model 2	
	Estimate	Std. Error	Estimate	Std. Error
Constant	2.01	1.05	1.03	.87
X_i	1.31	.67	2.36	.54
Z_i	-1.91	1.54		
$X_i \times Z_i$	2.00	0.95		
r^2	.664		.52	
$\hat{\sigma}$	1.41		1.59	

The data have the following descriptive statistics:

	Mean	Min	Max
X_i	1.47	0.28	2.51
Z_i	0.50	0.00	1.00
y_i	4.51	1.39	8.14

(a): (3 points) How many of the job applicants are white?

Answer: $.5 \times n = .5 \times 20 = 10$.

(b): (4 points) The model implies two relationships between preemployment test score and job performance, one for whites and one for minorities. What are the slope and intercept parameters of these separate relationships?

Answer: For whites, $Z_i = 0$, and so $E(y_i | X_i, Z_i = 0) = 2.01 + 1.31X_i$. For minorities, $Z_i = 1$ and so $E(y_i | X_i, Z_i = 1) = (2.01 - 1.91) + (1.31 + 2.00)X_i = .10 + 3.31X_i$.

(c): (5 points) How would you test whether the relationship between pre-employment test score and job performance differs across the two racial groups? Can you test that hypothesis with the information provided above? n.b., you don't actually have to compute the test.

Answer: *F*-test of the two implied restrictions (that the slope and intercept offsets across the groups are both non-zero). We actually do have an enough information to implement this test, from the two reported sets of regression results, using the two estimates of σ ; these provide enough information to compute the residual sums of squares used in *F* test statistic (i.e., $RSS = (\hat{\sigma})^2(n - k)$), which will have two degrees of freedom in the numerator, and $n - k$ (from the unrestricted model) or 16 degrees of freedom in the denominator.

Not necc for full credit: from lecture slides on *F*-testing

$$\frac{(RSS_r - RSS)/q}{RSS/(n - k)} = \frac{(RSS_r - RSS)/q}{\hat{\sigma}^2} \sim F_{q, n-k}$$

remembering that q is the number of linear restrictions being tested.

(d): (4 points) For this particular job performance assessment protocol, $y^* = 4$ will be used as a cut-off on hiring decisions (i.e., applicants with $y < 4$ are not hired). For both racial groups, compute the preemployment test score that yields the minimum acceptable job performance score. Comment briefly on the result.

Answer: For whites, $y^* = 4 = 2.01 + 1.31X_i \rightarrow X_i = (4 - 2.01)/1.31 = 1.52$. For minorities, $y^* = 4 = .10 + 3.31X_i \rightarrow (4 - .1)/3.31 = 1.17$. Note that the "critical" test score for minorities is substantially lower than for whites, and the difference of about .4 is large when considers that test scores only range between about 0 and 2.5.

Question 2: (7 points) What does it mean for an estimator to be consistent? If you can give a formal definition, be sure to also give an explanation in words that could be understood by a colleague who has not taken a class in statistics.

Answer: As sample size n tends to infinity, an estimator $\hat{\theta}_n$ converges on its value in the population θ . Formally, $\hat{\theta}_n$ is a **consistent** estimator of θ if

$$\lim_{n \rightarrow \infty} Pr(|\hat{\theta}_n - \theta| < \varepsilon) = 1$$

for any arbitrarily small positive ε . This definition of consistency is usually more compactly written as

$$\text{plim } \hat{\theta}_n = \theta.$$

Informally, as we throw more and more data at an estimation problem, the particular statistical procedure we are using (e.g., OLS, GLS, EGLS, MLE) gives

answers that “get closer” to the population parameter. This is a weak property of an estimator, and indeed, if a statistical procedure should do anything, it should do this.

Not necessary for full credit: Two conditions are sufficient for establishing the consistency of an estimator:

(a): asymptotic unbiasedness: $\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta$

(b): asymptotic variance is 0: $\lim_{n \rightarrow \infty} \text{var}(\hat{\theta}_n) = 0$.

Grading notes: partial credit for imprecise statement of intuition (full credit for clear statement of intuition or formal statement).

Question 3: (5 points) A researcher estimates a regression with $n = 100$ iid observations. The researcher comes to you with a methodological question: if she were to go out and collect 3 times as much data (which will also be iid) how much smaller would the standard errors of her regression be? What is your answer?

Answer: With independent data, statistical precision increases with the square root of sample size. So, on average (i.e., modulo sampling error), the research will obtain $\sqrt{3} \approx 1.73$ as much precision, or standard errors about $1/\sqrt{3} \approx 58\%$ the size of what the standard errors obtained with $n = 100$.

Grading notes: points of partial credit for knowing that the answer vaguely involves the square root of the sample size, three for having the correct formula for $\hat{\sigma}^2$ written down.

Question 4: Consider the data on y and x represented as a scatterplot in Figure 1.

(a): (5 points) Consider estimating the regression model $E(y) = \beta_0 + \beta_1 x$. Why are the estimates of β_0 and β_1 unlikely to be BLUE?

Answer: The residuals for the linear regression of y on x will be heteroskedastic; $\text{var}(u_i) \neq \sigma^2 \forall i$. OLS is not BLUE in this situation. The OLS estimates of the regression coefficients will have standard errors that are too large, relative to those of a BLUE estimator.

(b): (5 points) How might you recover BLUE estimates with these data?

Answer: GLS/EGLS. Re-run the regression, but with weights equal to $1/\sigma_i$. That is, create new variables $y_i^* = p_i y_i$ and $\mathbf{x}_i^* = p_i \cdot \mathbf{x}_i$ where $p_i = 1/\sigma_i$, and run the regression

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}^*$$

The disturbances from this regression are guaranteed to be homoskedastic. If σ_i is known, then this procedure is GLS; if instead σ_i must be estimated from the OLS residuals, then we have EGLS.

(c): (5 points) Now consider some additional information about these data. Specifically, Figure 2 shows the data from Figure 1 plotted by a third, binary variable z (solid squares indicate $z = 0$, and open squares $z = 1$). In light of this additional information, how would you re-analyze these data?

Answer: This additional information reveals the issue with these data to be structural shift, not heteroskedasticity. There are two regression regimes in the data: one for each value of z . If we had this information we would be better off running the regression of y on x conditioning on z , say with dummy variables and interaction terms:

$$y_i = \beta_0 + \alpha_0 z_i + x_i(\beta_1 + \alpha_1 z_i) + u_i$$

Question 5: (4 points) A model with a high r^2 but none or relatively few statistically significant coefficients is an indication of

- (a): omitted variable bias
- (b): heteroskedasticity
- (c): multicollinearity
- (d): non-normal disturbances

Answer: (c) multicollinearity.

Question 6: Consider the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$.

(a): (5 points) What are the *consequences* of violation of the assumption $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}) = \sigma^2\mathbf{I}_n$?

Answer: The OLS estimator is no longer “best” in the class of linear unbiased estimators. The estimated variances and covariances of $\hat{\boldsymbol{\beta}}$ are wrong, possibly leading to invalid inferences. Note also that the OLS estimator of σ^2 is biased in this situation.

(b): (10 points) Heteroskedasticity and autocorrelation are two different ways that this assumption can be violated. Describe how each of these phenomena can arise in social-science data.

Answer: Heteroscedasticity --- $\text{var}(u_i) = E(u_i^2) \neq \sigma^2, \forall i = 1, \dots, n$ --- this can arise in cross-sectional data when pooling disparate sets of observations (e.g., pooling over an income distribution), such that subsets of the data are more or less predictable than other subsets.

Autocorrelation: disturbances are not independent --- $E(u_i u_j) \neq 0, \forall i \neq j$ --- most commonly seen in time series data when adjacent observations have correlated disturbances, i.e., $\text{cor}(u_t u_{t-1}) = \rho$.

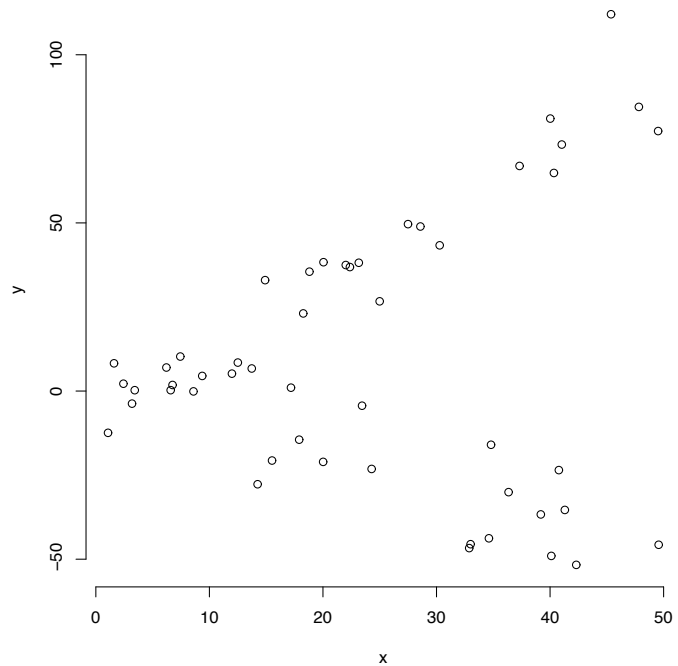


Figure 1: Scatterplot of Hypothetical Data.

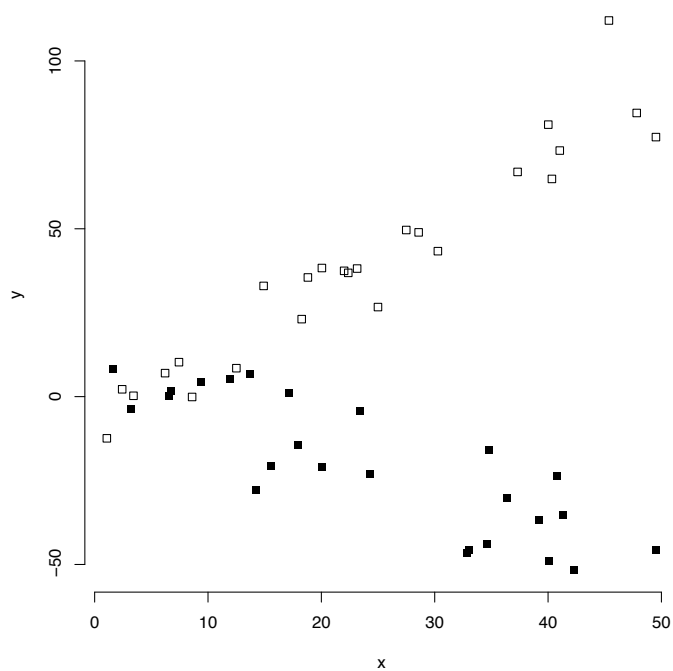


Figure 2: Scatterplot of Hypothetical Data.

(c): (5 points) Sketch a diagnostic residual plot typical of the pattern one would see with a moderate to high level of residual serial autocorrelation. Title and label your graph neatly and accurately.

Answer: Either a plot of the residuals over time, showing some kind of mean-crossing series, or a scatterplot of the residuals against lagged residuals, showing some kind of linear dependence between the two.

(d): (8 points) How would you test the suspicion that $\sigma_i^2 = \sigma^2 X_i^2$? What would you do if your suspicions were confirmed?

Answer: GLS if we *knew* (EGLS if we only *had good reasons to suspect*) that $\sigma_i^2 = \sigma^2 X_i^2$. Form weighted regressors $\mathbf{y}^* = \mathbf{P}\mathbf{y}$, $\mathbf{X}^* = \mathbf{P}\mathbf{X}$, where \mathbf{P} is a diagonal n by n matrix with $1/X_i$ on the leading diagonal.

Not necessary for full credit: Note the disturbance term in this regression is $\mathbf{u}^* = \mathbf{P}\mathbf{u}$ and so $E(\mathbf{u}^* \mathbf{u}^{*\prime}) = E(\mathbf{P}\mathbf{u}\mathbf{u}'\mathbf{P}') = \sigma^2 \mathbf{P}\mathbf{\Psi}\mathbf{P} = \sigma^2 \mathbf{I}$, since by definition $E(\mathbf{u}\mathbf{u}') = \sigma^2 \mathbf{\Psi}$ and by construction $\mathbf{\Psi} = \mathbf{P}^{-1}\mathbf{P}'^{-1}$. The GLS estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{y}^*$ will be BLUE.

(e): (6 points) A researcher says that he used heteroskedasticity-robust standard errors in presenting his results and in making inferences about $\boldsymbol{\beta}$. Explain how what the researcher did differs from OLS and EGLS (4 points each).

Answer: The use of heteroskedasticity-robust standard errors differs from OLS in that we use OLS to generate an estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, but we don't use the OLS estimate of the variance-covariance matrix $V(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$. Instead, we use $V(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\boldsymbol{\Omega}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ where typically $\tilde{\boldsymbol{\Omega}} = \text{diag}(\hat{\varepsilon}_i^2)$, where $\hat{\varepsilon}_i$ are the OLS residuals. This matrix is a consistent estimate of the variance-covariance matrix of the $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ in the presence of heteroskedasticity.

Contrast EGLS, where we use $\hat{\boldsymbol{\beta}}_{\text{EGLS}} = (\mathbf{X}'\hat{\boldsymbol{\Psi}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Psi}}^{-1}\mathbf{y}$, where $\hat{\boldsymbol{\Psi}}$ is an estimate of $\text{var}(\boldsymbol{\varepsilon}|\mathbf{X})$, and so is a different estimator than OLS (OLS with or without heteroskedasticity-robust standard errors).

Question 7: Consider the pattern of data shown in Figure 3. Each data point belongs to one of 10 groups. The plotted symbols on the graph, "1", "2", etc indicate the group membership of the corresponding data point.

(a): (2 points) Is $\hat{\beta}_1$ positive or negative?

Answer: n.b., β_1 here is a slope parameter in the model $E(\mathbf{y}|\mathbf{X}) = \beta_0 + \beta_1\mathbf{X}$. The estimate of this quantity given the data in the scatterplot will be positive.

(b): (4 points) If we estimated a 2nd regression model that included fixed effects for the 10 groups, what would happen to the estimate of β_1 ?

Answer: $\hat{\beta}_1$ will be negative in the presence of the fixed effects.

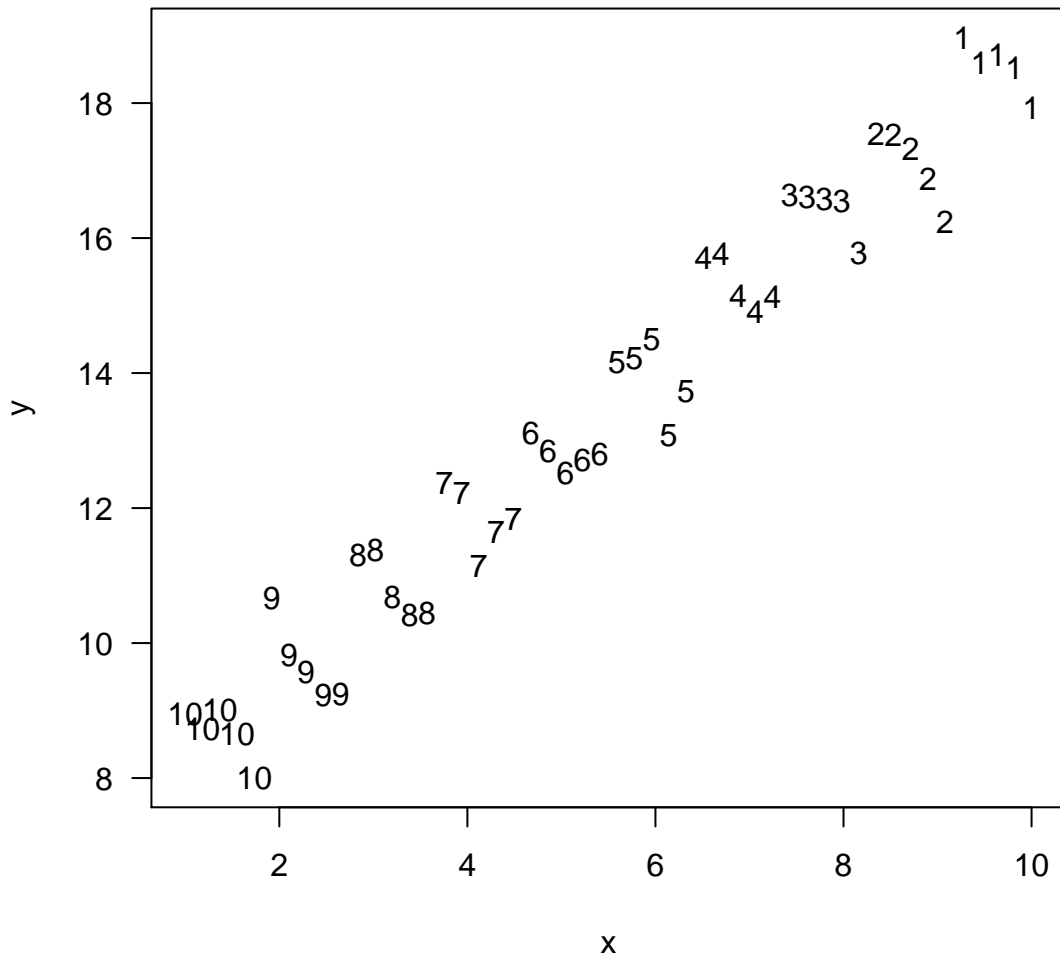


Figure 3: Scatterplot for Question 7

(c): (6 points) Based on your answer to the previous question, is the estimate of β_1 from the regression of y on x (shown in Figure 3) biased or unbiased?

Answer: Biased, in a classic omitted variables sense.

Not necessary for full credit: the fixed effects are clearly positively correlated with X . Hence the bias in the naïve estimate of β_1 is $E[(X'X)^{-1}X'\alpha]$ where α are the fixed effects; since $E(X'\alpha) > 0$ we have an overestimate of effect of X on y .

Question 8: (5 points) In time series data, what is the “spurious regression” problem and how does it arise?

Answer: If y_t is a lengthy, explosive time series and x_t is an equally lengthy explosive time series then the regression of y_t on x_t will usually result in a statistically significant t statistic on the slope coefficient, even though the two series may actually be independent.

Question 9: Suppose that, for a given state in the United States, you wish to use annual time series data to estimate the effect of the state-level minimum wage on the employment of those 18 to 25 years old (EMP). A simple model is

$$gEMP_t = \beta_0 + \beta_1 gMIN_t + \beta_2 gPOP_t + \beta_3 gGSP_t + \beta_4 gGDP_t + \varepsilon_t$$

where MIN_t is the minimum wage in real dollars, POP_t is the population from 18-25 years old, GSP_t is gross state product, and GDP_t is U.S. gross domestic product. The g prefix indicates the growth rate from year $t - 1$ to t .

(a): (5 points) If we are worried that the state chooses its minimum wage partly based on factors that affect youth employment, but that are unobserved (unobserved to us), what is the problem with OLS estimation?

Answer: In this case, we have reason to believe that $gMIN$ will be correlated with the disturbance term, and so the OLS estimates of β_1 will be biased and inconsistent (as will the estimates of the effects of any variable correlated with $gMIN$).

(b): (5 points) Let $USMIN_t$ be the U.S. minimum wage, which is also measured in real terms. Do you think $gUSMIN_t$ is uncorrelated with ε_t ?

Answer: Given that changes in the state’s minimum wage appears in the model (and so soaks up the part of $gEMP$ that responds to minimum wages), along with other indicators of state and federal economic activity, it seems plausible that changes in the federal minimum wage are uncorrelated with the residuals.

(c): (5 points) By law, any state’s minimum wage must be at least as large as the U.S. minimum wage. Explain why this makes $gUSMIN_t$ a potential IV candidate for $gMIN_t$.

Answer: The fact that any state's minimum wage must be at least as large as the U.S. minimum wage will induce a reasonable correlation between $gUSMIN$ and $gMIN$. Thus, $gUSMIN$ meets the criteria for a decent instrument: uncorrelated with disturbances in the equation of substantive interest, but a reasonable predictor of the endogenous regressor.

Question 10: Maximum likelihood.

(a): (3 points) What is a likelihood function? Provide as precise a definition as you can.

Answer: A likelihood function is the function that arises from expressing the joint probability density function (or mass function) of some data y , given a parameter θ as a function of θ . That is, $\mathcal{L}(\theta; y) = p(y; \theta)$.

(b): (5 points) Given the regression model $E(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$, under certain conditions it can be shown that the maximum likelihood estimator of $\boldsymbol{\beta}$ is the least squares estimator, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. What are those conditions? Be precise.

Answer: A sufficient condition $\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, with \mathbf{X} full rank. That is, we require (1) a normal distribution for \mathbf{y} conditional on \mathbf{X} ; (2) conditional independence, which is actually a weaker condition than the conditionally iid assumption $V(\mathbf{y}|\mathbf{X}) = \sigma^2\mathbf{I}_n$ given above, but iid will certainly do the trick; (3) that $(\mathbf{X}'\mathbf{X})^{-1}$ exist.

(c): (5 points) Provide an example where least squares and maximum likelihood yield different estimates. Under what conditions is this difference inconsequential?

Answer: In the regression model $\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, least squares estimates σ^2 as $\hat{\sigma}_{OLS}^2 = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}/(n - k)$ while the MLE is $\hat{\sigma}_{MLE}^2 = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}/n$. Similarly, when estimating a normal mean, we have $k = 1$ and the same result. There are other examples in the world of statistics, but these two are the ones I imagine you'd give as answers.

(d): (8 points) In a study of public opinion in France, Simon Jackman and Paul Sniderman measured survey respondents' general ideological position on a zero-to-one right-to-left scale (LibIssue), their level of knowledge and awareness about politics (PollInfo) on a zero-to-one scale (low to high) and whether respondents agree ($y = 1$) or disagree ($y = 0$) with the proposition that "more must be done to help the unemployed". The responses were analyzed using logit, yielding the following parameter estimates (obtained via maximum likelihood, standard errors in parentheses):

	Model 1	Model 2
Intercept	-1.62 (.17)	.29 (.49)
Liblssue	2.50 (.28)	-.52 (.81)
PollInfo		-3.74 (.93)
Liblssue × PollInfo		5.85 (1.48)
Log Likelihood	-1295.90	-1287.50

Test the restrictions of Model 1 vis-a-vis Model 2. Clearly state the null hypothesis, the test statistic, and the conclusion of the statistical test.

Answer: Model 1 nests inside Model 2 with two parameter restrictions. The null hypothesis is that the absent coefficients in Model 1 (the direct effect of PollInfo and the interaction of PollInfo with Liblssue) are both zero. Twice the difference in the log-likelihoods of the two models is distributed χ^2 with 2 degrees of freedom: twice this difference is 16.8 while the $p = .05$ critical value of the χ^2_2 distribution is 5.99, so we overwhelmingly reject the null hypothesis that restrictions of Model 1 are true, in favor of the richer specification of Model 2.

Question 11: (6 points) A researcher estimates a logit (or probit) model, $\Pr(y_i = 1 | \mathbf{x}_i) = F(\mathbf{x}_i \boldsymbol{\beta})$ where \mathbf{x}_i is a vector of independent variables, $i = 1, \dots, n$. However, the researcher does not include a “1” in the \mathbf{x}_i , suppressing the intercept term. What strong assumption is the researcher making in estimating a logit/probit model with this restriction? [Hint: what is the interpretation of the intercept in a regular regression model; translate that interpretation into the logit/probit context.]

Answer: Consider logit, $p_i = F(\mathbf{x}_i \boldsymbol{\beta})$ where F is the logistic CDF, and so

$$\ln \left(\frac{p_i}{1 - p_i} \right) = \mathbf{x}_i \boldsymbol{\beta} \quad \text{and so}$$

$$p_i = \frac{1}{1 + \exp(-\mathbf{x}_i \boldsymbol{\beta})}$$

The intercept insures that simply because all the predictors are set to zero, this does not necessarily imply that $\mathbf{x}_i \boldsymbol{\beta} = 0$ (i.e., without the intercept, the model “runs through the origin”). If there is no intercept then the we have the

logit/probit equivalent of the “model running through the origin”:

$$\begin{aligned} p_i | (\mathbf{x}_i = \mathbf{0}) &= F(0) \\ &= \frac{1}{1 + \exp(0)} \\ &= \frac{1}{1 + 1} \\ &= .5 \end{aligned}$$

This is far from innocuous, imposing the rather odd restriction that when all the predictors are set to 0, the model generates 50-50 predictions of the binary outcome. This constraint will rarely be sensible given the scores of the \mathbf{X} variables in any given data set, and will almost certainly seriously bias the coefficients on the included predictors (i.e., a particularly nasty form of omitted variable bias). We get the same result when we suppress the intercept in probit and for any binary response model using a F function with the property that $F(0) = .5$ (any CDF symmetric around zero has this property).

END OF EXAM

Total Number of Points: 140

df	Upper Tail Area				
	.25	.10	.05	.01	.001
2	2.77	4.61	5.99	9.21	13.8
3	4.11	6.25	7.81	11.3	16.3
4	5.39	7.78	9.49	13.3	18.5
5	6.63	9.24	11.1	15.1	20.5
6	7.84	10.6	12.6	16.8	22.5
7	9.04	12.0	14.1	18.5	24.3
8	10.2	13.4	15.5	20.1	26.1
9	11.4	14.7	16.9	21.7	27.9
10	12.5	16	18.3	23.2	29.6
11	13.7	17.3	19.7	24.7	31.3
12	14.8	18.5	21.0	26.2	32.9
13	16.0	19.8	22.4	27.7	34.5
14	17.1	21.1	23.7	29.1	36.1
15	18.2	22.3	25	30.6	37.7
20	23.8	28.4	31.4	37.6	45.3
30	34.8	40.3	43.8	50.9	59.7
50	56.3	63.2	67.5	76.2	86.7
100	109	118	124	136	149
200	213	226	234	249	268
300	316	332	341	360	381
500	521	541	553	576	603
1000	1030	1058	1075	1107	1144
3000	3052	3100	3129	3183	3245

Table 1: Critical values of the χ^2 distribution.