

Question 1: (5 points) After estimating the least squares regression of \mathbf{y} on \mathbf{X} , a researcher finds that the correlation between the regression's predicted values $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ and the residuals $\mathbf{y} - \hat{\mathbf{y}}$ is approximately zero. What can the researcher conclude? [Hint: no proof necessary].

Answer: Nothing. $E(\hat{\mathbf{y}}'\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}$, by construction, for least squares fits. That is, irrespective of whether any assumptions about the disturbances etc etc do or do not hold, this correlation will always be zero, up to errors induced by floating point representation. Hence this correlation is not informative about lack of fit, appropriateness of assumptions, etc.

Proof, not necessary for answer:

$$\begin{aligned}\hat{\mathbf{y}}'\hat{\boldsymbol{\varepsilon}} &= (\mathbf{H}\mathbf{y})'(\mathbf{M}\mathbf{y}) \\ &= \mathbf{y}'\mathbf{H}'\mathbf{M}\mathbf{y} \\ &= \mathbf{y}'\mathbf{H}'(\mathbf{I} - \mathbf{H})\mathbf{y} \\ &= \mathbf{y}'\mathbf{H}'\mathbf{y} - \mathbf{y}'\mathbf{H}'\mathbf{H}\mathbf{y} \\ &= \mathbf{0},\end{aligned}$$

recalling that $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}' = \mathbf{H}'\mathbf{H}$ (i.e., \mathbf{H} is a symmetric idempotent matrix).

Grading Note: to receive full credit you needed to write that “ $E(\hat{\mathbf{y}}'\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}$ by construction.” Otherwise one point was taken off.

Question 2: For each of the following questions, simply select the correct response. No explanation is necessary.

(a): (4 points) A regression that has many statistically insignificant predictors, but a relatively high r^2 is most likely to be suffering from...

- (a) group-wise heteroskedasticity
- (b) data points with high leverages
- (c) data points with high influence
- (d) data that are skewed right
- (e) multicollinearity

Answer: (e).

(b): (4 points) We would use a Chow test to test for...

- (a) heteroskedastic disturbances
- (b) exogeneity of instruments
- (c) structural stability in regression coefficients
- (d) endogeneity of regressors

(e) autocorrelated disturbances

Answer: (c).

(c): (4 points) We would use a Durbin-Wu-Hausman test to test for...

- (a) heteroskedastic disturbances
- (b) exogeneity of instruments
- (c) structural stability in regression coefficients
- (d) endogeneity of regressors
- (e) autocorrelated disturbances

Answer: (b)

(d): (4 points) We would use a Breusch-Pagan test to test for...

- (a) heteroskedastic disturbances
- (b) exogeneity of instruments
- (c) omitted variable bias
- (d) endogeneity of regressors
- (e) autocorrelated disturbances

Answer: (a)

(e): (4 points) A researcher reporting the results of a regression analysis says that they have used a “HCCM”. This is because the researcher is concerned about...

- (a) regressors measured with error
- (b) heteroskedastic disturbances
- (c) exogeneity of instruments
- (d) omitted variable bias
- (e) endogeneity of regressors

Answer: (b)

(f): (4 points) A researcher conducts a regression analysis with all **X** variables centered around their respective averages. This implies that...

- (a) all the estimated slope coefficients in the regression model will be equal to 1.0
- (b) that the estimated intercept will be equal to the mean of y .
- (c) the intercept in the regression equation will be zero, and so can be dropped from the model
- (d) the estimated coefficients will have smaller standard errors than if the model was estimated with the uncentered variables
- (e) none of the above

Answer: (b)

(g): (4 points) Complete the 2nd sentence. A researcher regresses y on a single predictor, X , and a constant. The r^2 from this regression is equal to...

Answer: The square of the correlation between y and X .

Grading Note: 3 points if you said that it was the variance in y explained by the variance in X . 2 points were given if you knew it was the correlation of y and X (but not the square).

(h): (4 points) Complete the sentence: “An estimator $\hat{\theta}$ whose sampling distribution (un-normalized by sample size) is asymptotically degenerate with point mass on θ is ...”

Answer: consistent.

(i): (4 points) A researcher runs the regression of a survey-based measure of public opinion in region $i = 1, \dots, n$, on various demographic characteristics of those regions. The surveys conducted in each region vary considerably in terms of the sample sizes that are used. The resulting regression is guaranteed to have...

- (a) spatially correlated disturbances
- (b) regressors measured with error
- (c) highly influential observations
- (d) heteroskedastic disturbances

Answer: (d)

(j): (4 points) In the presence of disturbances that are not “iid”, the OLS estimator of β is generally...

- (a) unbiased and consistent
- (b) the best linear unbiased estimator
- (c) biased but consistent
- (d) inconsistent

Answer: (a).

Question 3: In the following question, consider data y_{ij} and predictors \mathbf{x}_{ij} where $j = 1, \dots, m_i$ indexes observations in geographic regions $i = 1, \dots, n$. The researcher estimates a regression of y_{ij} on predictors \mathbf{x}_{ij} , including a vector of binary indicator (or “dummy”) variables \mathbf{d}_{ij} , each one coded 1 if the particular observation is from region i and coded 0 otherwise.

(a): (2 points) How many dummy variables appear in the estimated regression? [Hint: there is more than one correct answer here; be clear].

Answer: n if an intercept is dropped from the model; $n - 1$ if an intercept is included in the model.

(b): (3 points) What name is conventionally given to the *coefficients* that attach to the region-specific dummy variables?

Answer: “fixed effects”.

(c): (3 points) When we estimate a model such as the one given above, what property do the residuals have, region-by-region?

Answer: The residuals have mean zero within each region.

(d): (4 points) The presence of these dummy variables in the model helps guard against what possible problem in the model specification? Hint: what does a regression look like without the region-by-region dummy variables, versus the regression that includes the **d** dummy variables?

Answer: The dummy variables for region pick up time-averaged, region-specific effects that are otherwise omitted from the model. If these effects are correlated with predictors **X** that do appear in the model, then omission of the region-specific dummies leads to omitted variable bias.

Grading Note: 3 points were given if you just said omitted variable bias without elaborating.

Question 4: A researcher conducting a regression analysis believes that at least one of her regressors is endogenous, and seeks to remedy the situation using instrumental variables. She seeks your advice regarding the following issues. Answer merely by selecting the correct response.

(a): (4 points) In the presence of an endogenous regressor, the least squares estimates of the parameters of the structural equation of substantive interest are...

- (a) biased, but consistent
- (b) unbiased and consistent
- (c) biased and inconsistent
- (d) unbiased, but inconsistent

Answer: (c)

(b): (4 points) Complete the sentence. The instrumental variables estimator of the parameters in the structural equation of substantive interest is just identified when...

Answer: There are as many excluded exogenous variables from the structural equation as there are endogenous regressors.

Grading Note: 3 points were given for saying that there are as many instruments as there are endogenous regressors.

(c): (4 points) Let the structural model of substantive interest be $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, where \mathbf{X}_1 are endogenous regressors and \mathbf{X}_2 are exogenous regressors. Let \mathbf{Z} be a matrix of exogenous regressors, with $\mathbf{X}_2 \notin \mathbf{Z}$. Then the “first stage” of the two-stage least squares estimator consists of regressing each endogenous regressor \mathbf{x}_j , where j indexes the columns of \mathbf{X}_1 , on...

- (a) \mathbf{Z}
- (b) \mathbf{X}_2 and \mathbf{Z}
- (c) \mathbf{X}_2 , \mathbf{Z} and $\mathbf{x}_k, \forall k \neq j$.
- (d) any of the above.

Answer: (b).

(d): (4 points) Using the notation in the previous question, the guiding principles for choosing \mathbf{Z} variables are...

- (a) $\text{plim}_{\frac{1}{n}}(\mathbf{Z}'\boldsymbol{\epsilon}) = \mathbf{0}$.
- (b) $\text{plim}_{\frac{1}{n}}(\mathbf{Z}'\mathbf{X}_1)$ be large.
- (c) in order of importance, (a), then (b)
- (d) in order of importance, (b), then (a)

Answer: (c).

(e): (4 points) Again using the notation from above, the weakness of an instrument (or collection of instruments) is best assessed by examining...

- (a) the correlations between \mathbf{y} and the columns of \mathbf{Z}
- (b) the correlations between the columns of \mathbf{X}_1 and the columns of \mathbf{Z}
- (c) the correlations between the columns of \mathbf{X}_2 and the columns of \mathbf{Z}
- (d) the partial correlations between the columns of \mathbf{X}_1 and the columns of \mathbf{Z} controlling for \mathbf{X}_2
- (e) any of (b), (c), or (d)
- (f) none of the above

Answer: (d).

Question 5: Professor Shanto Iyengar and I (Jackman) have data assessing reactions to political advertising on television. Each subject ($i = 1, \dots, n$) is shown a 30 second ad, and is given a “dial”, a device that the respondent turns to the left or the right to indicate “disliking” or “liking” what they are viewing as the ad progresses. The dial generates readings Y_{it} on a 0 - 100 scale every second ($t = 0, \dots, 30$), which are captured and stored on a computer. The dial is initialized at $Y_{i0} = 50 \forall i$. Iyengar and Jackman considered fitting the model

$$E(Y_{it}|t) = \alpha \exp[-\gamma \exp(-\beta t)]$$

to the data, where α , γ and β are unknown parameters to be estimated.

(a): (5 points) The dials data are constrained to start at 50 at $t = 0, \forall i$ and to lie in the 0-100 interval. What might this imply about the disturbances $Y_{it} - E(Y_{it}|t)$ of the model given above?

Answer: The data display no variation at all at $t = 0$, but display variation later in time, so almost surely, these data are heteroskedastic.

(b): (5 points) Since $Y_{it}(t = 0) = 50 \forall i$, what constraint is implied on one or more of the parameters in the model?

Answer: At $t = 0$, $\exp(-\beta t) = \exp(0) = 1, \forall \beta$. But since we have $Y_{it}(t = 0) = 50$, this means that

$$\begin{aligned} 50 &= \alpha \exp[-\gamma], \Rightarrow \\ \log(50) &= \log(\alpha) - \gamma, \Rightarrow \\ \gamma &= \log(\alpha/50) \end{aligned}$$

(c): (5 points) Irrespective of the answers to the previous questions, can ordinary least squares be used to estimate the parameters of the proposed model? Explain your answer.

Answer: No. There is no way to transform the equation so as to make the model linear in the parameters (or at least none that I can see).

Question 6: Suppose we have data (y_i, \mathbf{x}_i) , $i = 1, \dots, n$, for which we posit the probability model $y_i | \mathbf{x}_i \stackrel{\text{iid}}{\sim} N(\mathbf{x}_i \boldsymbol{\beta}, \sigma^2)$, where $\boldsymbol{\beta}$ is a vector of unknown parameters and σ^2 is an unknown scalar.

(a): (3 points) What is the maximum likelihood estimate of $\boldsymbol{\beta}$?

Answer: $\hat{\boldsymbol{\beta}}_{\text{MLE}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

(b): (3 points) What is the maximum likelihood estimate of σ^2 ?

Answer:

$$\hat{\sigma}_{\text{MLE}}^2 = \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n} = \frac{\mathbf{y}'\mathbf{M}\mathbf{y}}{n}$$

where $\mathbf{M} = \mathbf{I} - \mathbf{H}$, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Grading note: some definition of $\hat{\boldsymbol{\epsilon}}$ in the numerator of the estimator is required, not necessarily the one I gave, but something reasonable.

Question 7: Let $\theta \in [0, 1]$ be the probability that a coin comes up “heads” when flipped. Suppose the coin is flipped five times and comes up heads each time. You can assume the outcomes of the coin flips are independent events.

(a): (3 points) What is the maximum likelihood estimate of θ ?

Answer: $\hat{\theta} = 1.0$. Proof: under independence, the likelihood function for these data is simply the product of the probabilities of the individual events. The probability of a “head” (H) is θ . Hence the probability of 5 heads is θ^5 . Since $\theta \in [0, 1]$ (i.e., θ is a probability), the likelihood is maximized by setting $\hat{\theta} = 1.0$.

(b): (5 points) What is the value of the likelihood function for these data evaluated at the maximum likelihood estimate of θ ?

Answer: $\mathcal{L}(\{H, H, H, H, H\}; \theta = \hat{\theta}_{\text{MLE}} = 1) = \prod_{i=1}^5 \theta = \theta^5 = 1^5 = 1$.

(c): (5 points) What is the value of the likelihood function for these data evaluated at $\theta = .5$?

Answer: $\mathcal{L}(H^5; \theta = .5) = \prod_{i=1}^5 \theta = .5^5 = .03125$.

(d): (7 points) Given the data, assess the plausibility of the hypothesis $H_0 : \theta = .5$ (the coin is “fair”) against the one-sided alternative $H_A : \theta > .5$. Be explicit as to how you conduct this test (there is more than one way to do it). What do you conclude?

Answer: Under $H_0 : \theta = .5$, the probability of 5 heads in independent tosses is just .03125, and so if we were using a conventional $\alpha = .05$ test, we would reject the null hypothesis in favor of the one-sided alternative (which is in the direction of the observed outcome of 5 heads).

A likelihood ratio test is also another way to do this. We might consider the null $H_0 : \theta = .5$ as a restriction to test, relative to the MLE of $\hat{\theta} = 1$. Note that in a large sample (!), twice the difference in the log-likelihoods from two models for the data is distributed χ^2 with 1 degree of freedom. The restricted log-likelihood under H_0 is $\log(.03125) = -3.47$ while the unrestricted log-likelihood is $\log(1) = 0$. Twice the difference is 6.93, which puts us far out into the tail of the χ^2_1 density; just .008 of the probability mass under the χ^2 density lies to the right of 6.93. That is, we overwhelmingly reject the restrictions in favor of the unrestricted model.

Grading note: Note that this likelihood ratio test isn't testing the one-sided alternative per se, but a specific alternative to H_0 .

Question 8: (6 points) Logistic regression models (i.e., logit models for binary data) are sometimes said to be non-linear models. What is linear about a logistic regression model, and what is non-linear? Be brief, but precise and explicit in your answer.

Answer: Logit models are non-linear in the sense that the mapping from the covariates to probabilities is non-linear; specifically,

$$\hat{P}(y_i = 1 | \mathbf{x}_i, \hat{\boldsymbol{\beta}}) = \frac{\exp(\mathbf{x}_i \hat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{x}_i \hat{\boldsymbol{\beta}})} = \frac{1}{1 + \exp(-\mathbf{x}_i \hat{\boldsymbol{\beta}})},$$

and so the impact of a unit change in x is not constant, but varies over x . Digression: if we re-write the equation above as $P_i = F(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$, then by the chain rule of differential calculus

$$\frac{\partial F(\mathbf{x}_i; \hat{\boldsymbol{\beta}})}{\partial \mathbf{x}_i} = f(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) \cdot \hat{\boldsymbol{\beta}}$$

where f is the logistic density. This makes it obvious that increases in x produce impacts on the probability of the binary outcome that vary over x , and why x variables have their biggest “bang-for-the-buck” when $p_i = .5$, because $f(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$ is its maximum there (when $\mathbf{x}_i \hat{\boldsymbol{\beta}} = 0$).

But logit is a linear model in the logits: that is,

$$E(\ln[(p_i/(1 - p_i))]) = y_i^* = \mathbf{x}_i \boldsymbol{\beta},$$

and so logit is “linear in the log-odds ratio” (e.g., a one unit increase in x produces $\boldsymbol{\beta}$ change in the log-odds-ratio y_i^*).

Question 9: On April 15th, 1912, the *Titanic* collided with an iceberg and sank with much loss of life. Logistic regression models were used to analyze data available for 2,201 passengers and crew: a binary dependent variable (coded 1 for survival, 0 otherwise), with predictors measuring class of travel (0 for crew, 1 for first class, 2 for second class, 3 for third class), adult/child, and gender. Table 1 summarizes this analysis, presenting maximum likelihood estimates of coefficients (with standard errors in parentheses) and empty table entries indicating whether that the corresponding variable was omitted from the respective logit model.

	Model 1	Model 2	Model 3	Model 4	Model 5
Intercept	-.74 (.05)	2.61 (.29)			1.27 (.34)
Class (0-3)		-.33 (.05)			
1st class (0-1)			3.11 (.30)	2.07 (.35)	.80 (.16)
2nd class (0-1)			2.09 (.28)	1.04 (.34)	-.23 (.18)
3rd class (0-1)			1.33 (.25)	.26 (.32)	-1.01 (.15)
Crew (0-1)			2.25 (.30)	1.27 (.34)	
Adult (0-1)		-1.01 (.25)	-1.06 (.24)	.11 (.34)	.11 (.34)
Male (0-1)		-2.61 (.13)	-2.42 (.14)	-.72 (.41)	-.72 (.41)
Adult × Male				-1.90 (.43)	-1.90 (.43)
Log-Likelihood	-1384.73	-1137.45	-1105.03	-1096.02	-1096.02

Table 1: Logit Estimates, Analysis of *Titanic* Disaster. Maximum likelihood estimates and standard errors in parentheses.

- (a): (5 points) What proportion of those on board survived the disaster?
Answer: Model 1 simply includes an intercept, and so the predicted probability from this model will be equal to the sample proportion of survivors. In this case, $\Pr(y_i = 1) = F(-.74) = .32$.
- (b): (5 points) What is the substantive content of the null hypothesis tested by a comparison of Models 2 and 3?
Answer: In model 2, class of passage enters as a continuous variable, scored zero through 3. In model 3, class of passage enters as a series of mutually exclusive and exhaustive dummy variables. Model 2 is thus a restrictive version of model 3, with the restriction being the linearity in changes in the log-odds ratio across the different classes of passage.
- (c): (5 points) Use a likelihood ratio test to compare Models 2 and 3.
Answer: Class of passage enters model 3 in a series of mutually exclusive and exhaustive dummy variables: 4 in all. In model 2, class of passage consumes just one degree of freedom. Thus, model 3 consumes 2 more degrees of freedom relative to model 2, recalling that the intercept of model 2 has been absorbed into the mutually exclusive and exhaustive dummy variables.
 Twice the difference in the log-likelihoods is $2 \times (1137.45 - 1105.03) = 2 \times 32.42 = 64.84$ which is overwhelmingly statistically significant (the p -value is roughly 8×10^{-15}).
- (d): (5 points) How do Models 4 and Model 5 differ? That is, what is being tested in Model 5, versus Model 4 (or vice-versa)?
Answer: In model 5, the dummy variable for “crew” has been absorbed into the intercept, meaning that the coefficient on the other class dummies can be interpreted as offsets relative to the “crew” baseline category. In this way we can test for differences against the baseline of “crew”. This wasn’t possible given the parameterization of Model 4.
- (e): (5 points) You need not compute a likelihood ratio test to compare Model 3 and 4. Why?
Answer: The only difference between Model 3 and Model 4 is the addition of the adult-male interaction variable. In a large sample, the z -statistic formed by dividing the MLE of the coefficient for that variable by its standard error will yield the same conclusions as the χ^2 test for the likelihood ratio test. We have a big sample here, so we really have all we need to test the statistical significance of that coefficient.
- (f): (5 points) A time-honored policy in disasters at sea is that “women and children” are the first to be rescued. How might we augment or change the analysis presented above to test the extent to which different classes adhered to the “women and children first” policy?
Answer: We’d want higher-order interaction terms, estimating an adult-male term for each class of passage. We could then compare this model with the class-specific adult-male terms against Model 5, via a likelihood ratio test.