STAT 608 - Final Exam
May 6, 2022

PART I: Multiple Choice (4 Points Per Question). Unless otherwise instructed, choose the **best** answer.

1. A linear regression model $y = \beta_1 x_1 + \beta_2 x_2 + e$ is fit to a dataset, and the VIF for x_1 is found to be 4. What does this mean?
   (a) **The variance of $\hat{\beta}_1$ is 4 times as large as it would have been if the predictors were independent.
   (b) The variance of $\hat{\beta}_1$ is 4 times as large as it would have been if the predictors were normally distributed.
   (c) The variance of x_1 is 4 times as large as the variance of x_2.
   (d) The variance of $\hat{\beta}_1$ is 4 times as large as the variance of $\hat{\beta}_2$.

2. Suppose the odds of an event occurring are greater than 1. What does that tell you about the probability of the event? (Be as specific as possible.)
   (a) The probability is less than 1.
   (b) The probability is less than 0.5.
   (c) **The probability is greater than 0.5.
   (d) The probability is less than 0.75.
   (e) The probability is greater than 0.75.

3. For a regression model with $n = 4$, it is known that the variance of the first two errors is σ^2, while the variance of the last two errors is $\sigma^2/4$. Which of the following are the appropriate choice for the weights in a weighted regression model?
   (a) $w_1 = 1$, $w_2 = 1$, $w_3 = 2$, $w_4 = 2$
   (b) $w_1 = 1$, $w_2 = 1$, $w_3 = 0.5$, $w_4 = 0.5$
   (c) **$w_1 = 1$, $w_2 = 1$, $w_3 = 4$, $w_4 = 4$
   (d) $w_1 = 1$, $w_2 = 1$, $w_3 = 0.25$, $w_4 = 0.25$

4. Consider a logistic regression model with a predictor variable x_1 that is right-skewed according to the gamma distribution. The log odds of a binary response variable could then be considered to be
   (a) A linear function of x_1. IF x_1 WERE POISSON OR BERNOULLI, THIS WOULD BE TRUE.
   (b) A function of x_1 and x_1^2.
   (c) **A function of x_1 and $\log(x_1)$.
   (d) A function of x_1, $\log(x_1)$, and x_1^2.

5. The admissions committee for TAMU has decided on the following model for admitting most of its students: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \texttt{HSGPA}_i + \hat{\beta}_2 \texttt{SAT}_i + \hat{\beta}_3 \texttt{iSports}_i$, where the response variable is the student's college GPA; `HSGPA` is the student's high school GPA; `SAT` is the student's SAT score; and `iSports` is an indicator of whether the student played sports in high school. As long as students' predicted GPA is higher than 2.5, students are admitted. Which of the following is the geometric interpretation of this model?

   (a) Two parallel lines
   (b) Two lines that cross
   (c) Two parabolas
   (d) **Two parallel planes (WE HAVE TWO QUANTITATIVE VARIABLES AND ONE CATEGORICAL VARIABLE WITH TWO CATEGORIES.)
   (e) Two planes that intersect

6. The same admissions committee from the previous problem is interested in fitting an interaction between `HSGPA` and `SAT`. What will this imply?

   (a) High school GPA and SAT score are substantially correlated with each other. (THIS IS MULTICOLLINEARITY, NOT INTERACTION.)
   (b) The slope for `HSGPA` and the slope for `SAT` are different. (HSGPA AND SAT ARE IN DIFFERENT UNITS, SO THEY WILL DEFINITELY HAVE DIFFERENT SLOPES.)
   (c) When SAT scores are added to the model, the slope for high school GPA changes.
   (d) **The relationship between high school GPA and college GPA differs depending on the value of SAT scores.

7. Which of the following plots should appear to be a random scatter with no pattern (not even a linear pattern) when all model assumptions are met for a linear regression model?

   (a) **A plot of residuals against a predictor variable. (REMEMBER THE RESIDUALS SHOULD BE INDEPENDENT (ORTHOGONAL) OF THE PREDICTORS AND THE FITTED VALUES.)
   (b) A plot of the response variable against a predictor variable.
   (c) A QQ plot of residuals.
   (d) A plot of predicted (fitted) values against the response variable.
   (e) A plot of predicted (fitted) values against a predictor variable.

8. Consider a study of the effect of a dietary supplement on cholesterol (`Chol`) in rabbits. Suppose that four rabbits were given 1, 2, 3, and 4 mg (`Dose`) of the supplement, and an indicator variable for whether each rabbit was male (`iMale`) was another predictor variable (no, no, yes, yes). Given the design matrix below, what did the model look like?

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 3 & 1 & 3 \\ 1 & 4 & 1 & 4 \end{bmatrix}$$

(a) $\texttt{Chol} = \beta_0 + \beta_1\texttt{Dose} + \beta_2\texttt{iMale} + e$

(b) $\texttt{Chol} = \beta_0 + \beta_1\texttt{Dose} + \beta_2\texttt{iMale} + \beta_3\texttt{Dose}^2 + e$

(c) $\texttt{Chol} = \beta_0 + \beta_1\texttt{Dose} + \beta_2\texttt{iMale} + \beta_3 \log(\texttt{Dose}) + e$

(d) **$\texttt{Chol} = \beta_0 + \beta_1\texttt{Dose} + \beta_2\texttt{iMale} + \beta_3 (\texttt{Dose} \times \texttt{iMale}) + e$

9. In the same study of rabbits as the previous problem, it was desired to predict $y =$ cholesterol using `Dose` and `iMale`. A 95% prediction interval for male rabbits who received 2.5 mg of the supplement was found to be $(141.5, 162.0)$. What does this mean?

   (a) We are 95% sure that the population slope for the model is between 141.5 and 162.0.

   (b) We are 95% sure the mean cholesterol for male rabbits who receive 2.5 mg of supplement is between 141.5 and 162.0. (WE'RE DOING PREDICTION, NOT A CONFIDENCE INTERVAL FOR MEAN RESPONSE.)

   (c) If we randomly select a male rabbit from the population of all male rabbits, there would be a 95% chance of selecting a rabbit with a cholesterol level between 141.5 and 162.0. (WE NEED ONLY AT $x = 2.5$ MG.)

   (d) **If we randomly select a male rabbit from the population of all male rabbits who received a 2.5 mg dose of supplement, there would be a 95% chance of selecting a male rabbit with cholesterol between 141.5 and 162.0.

   (e) If we randomly select a group of male rabbits from the population of all male students who received a 2.5 mg dose of supplement, there would be a 95% chance of selecting a sample mean cholesterol between 141.5 and 162.0. (WE'RE DOING PREDICTION, NOT A CONFIDENCE INTERVAL FOR MEAN RESPONSE.)

10. As teenagers grow older, their sleep habits change. The data summarized below are from a random sample of 346 teens; researchers are particularly interested in modeling the number who get at least 7 hours of sleep, using age as a predictor. Which of the following is true about the model?

Sleep	Age					
At Least 7 Hours	14	15	16	17	18	Total
No	12	35	37	39	27	150
Yes	34	79	77	65	41	296
Total	46	114	114	104	68	446

   (a) For each age, the number getting at least 7 hours is possibly Poisson, so we might take the square root of the response.

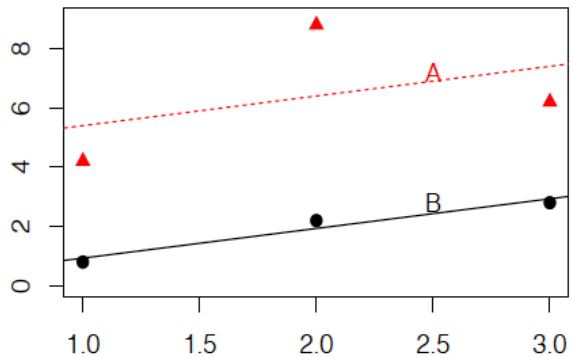
   (b) **For each age, the number getting at least 7 hours is possibly binomial, so we might use logistic regression. (WE WANT TO MODEL THE PROPORTION WHO GET AT LEAST 7 HOURS OF SLEEP FOR EACH AGE.)

   (c) Age and the number getting at least 7 hours are both skewed, so a log/log transformation seems best.

   (d) Age is uniformly distributed, so we should use both age and its square in the model predicting the number getting at least 7 hours.

Part II: Multiple Select. For each problem below, you may choose more than one answer; circle the letters corresponding to all that apply.

11. Below are plotted to scale two generated datasets, both having 3 points with the x-values the same. Simple linear regression models have been fitted to both datasets; the dotted line (Model A) is fitted to the triangle points, while the solid line (Model B) is fitted to the circle points. The slopes in both models are equal to 1. Which of the following is / are true? Select all that apply. (4 points)



   (a) SSTot for Model A is equal to SSTot for Model B.
   (b) **SSTot for Model A is larger than SSTot for Model B. (Because the y values are more variable.)
   (c) SSTot for Model A is smaller than SSTot for Model B.
   (d) RSS for Model A is smaller than RSS for Model B.
   (e) **RSS for Model A is larger than RSS for Model B. (The points are closer to the line for model B.)
   (f) RSS for Model A is equal to RSS for Model B.

12. Suppose a simple linear regression model has a single "good" leverage point: it doesn't fall far away from the general pattern of increasing y as x increases that the other points seem to follow. Which of the following will be substantially affected if we remove this point from the dataset? Select all that apply. (4 points)

   (a) **Correlation between x and y (Correlation is not robust to outliers: notice the numerator involves $(x_i - \bar{x})(y_i - \bar{y})$, and the good leverage point will contribute an outlying value to that computation.)
   (b) Sample slope (It's "bad" leverage points that change slope.)
   (c) **P-value for slope (Outliers still affect inference.)

13. Which of the following must be met in order for a multiple regression model (assume no transformations are necessary) to be valid? Select all that apply. (6 points)

   (a) The R^2 must be high.

(b) The errors must be normally distributed. (This is necessary for prediction intervals and small sample sizes based on normality assumptions, but the model can still be valid if this is not met, and we can still bootstrap any inference procedures from the distribution of the residuals.)

(c) **The predictors must be reasonably close to mutually independent.

(d) **The mean function must be modeled correctly.

(e) **The errors must have constant variance.

Part III: Long Answer. Explain all answers as if to a layman whenever possible, and show all work.

14. We have data on crime rates (number of offenses per million population) (`Y`) in terms of the following explanatory variables:

- `Youth`: number of males aged 18-24 per 1000
- `Education`: average number of years of schooling
- `ExpenditureYear0`: per capita expenditure on police
- `LabourForce`: number per 1000 of males aged 18-24 who are employed
- `StateSize`: state size in hundred thousands
- `Wage`: median weekly wage

(a) The results of fitting the full model:

$$\texttt{Y}_i = \beta_0 + \beta_1\texttt{Youth}_i + \beta_2\texttt{Education}_i + \beta_3\texttt{ExpenditureYear0}_i + \beta_4\texttt{LabourForce}_i + \beta_5\texttt{StateSize}_i + \beta_6\texttt{Wage}_i + e_i$$

and the reduced model:

$$\texttt{Y}_i = \beta_0 + \beta_1\texttt{Youth}_i + \beta_2\texttt{ExpenditureYear0}_i + e_i$$

are shown in the Appendix. Perform a hypothesis test for whether the reduced model is a significantly poorer fit than the full model. Report a p-value and interpret the results. A few probability values are shown at the end of the Appendix.(8 points)

The relevant F statistic is

$$F = \frac{(\texttt{RSS}_{\text{RED}} - \texttt{RSS}_{\text{FULL}})/(\texttt{df}_{\text{RED}} - \texttt{df}_{\text{FULL}})}{\texttt{RSS}_{\text{FULL}}/(n-p-1)} = \frac{(20.53^2 * 44 - 21.15^2 * 40)/4}{21.15^2} = 0.3645$$

The p-value in R is `1 - pf(0.3645, 4, 40)` $= 0.8324$. Since this is greater than $\alpha = 0.05$, we fail to reject the null hypothesis that $\beta_2 = \beta_4 = \beta_5 = \beta_6 = 0$ and conclude that the reduced model is sufficient.

(b) Figure 1 in the Appendix shows a pairs plot of the crime data. Also shown is the sample correlation matrix and VIF values for each explanatory variable. Does there appear to be any problems with multicollinearity? Why or why not? (5 points)

There is moderate evidence of multicollinearity. For `Wage`, in particular, there are a few scatterplots with other explanatory variables

THAT SHOW RELATIVELY STRONG LINEAR ASSOCIATIONS. THESE ARE QUANTIFIED WITH MODERATE SAMPLE CORRELATIONS, WITH THE MAXIMUM BEING WITH VARIABLE `ExpenditureYear0`. THE VIF FOR `Wage` IS ABOUT 4.5, WHICH IS CLOSE TO OUR RULE OF THUMB THRESHOLD OF 5.0.

(c) The result of calling `regsubsets` on the full model is in the Appendix. Table 1 shows BIC values for models with different numbers of explanatory variables. According to the BIC criterion, which model should we choose, and why? (5 points)

THE SMALLEST BIC VALUE IS 292.53, CORRESPONDING TO THE MODEL WITH `Youth` AND `ExpenditureYear0` IN IT.

(d) Figure 2 in the Appendix shows a plot of regression coefficients according to the lasso. Each line is labeled with a number, and that number indicates the explanatory variable number as ordered at the beginning of this problem (1 corresponds to `Youth`, 2 to `Education`, etc.). Consider the lasso model with the same number of explanatory variables as the model chosen by BIC. Are the same explanatory variables selected by lasso, and why or why not (based just on the figure)? (5 points)

YES, THE SAME VARIABLES ARE SELECTED FOR THE TWO-VARIABLE MODEL. YOU CAN SEE THIS BY THE CURVES FOR VARIABLES 1 AND 3 BEING ABOVE 0 WHEN ALL OTHERS ARE EQUAL TO 0. THIS MEANS THAT ONLY THE COEFFICIENTS FOR VARIABLES 1 AND 3 (`Youth` AND `ExpenditureYear0`) ARE NOT EQUAL TO 0 IN THE TWO-VARIABLE MODEL.

15. Now consider the binary indicator variable for whether $\texttt{Y} > 100$.

(a) The results of fitting the full model:

$$\log\left(\frac{\theta}{1-\theta}\right) = \beta_0 + \beta_1\texttt{Youth}_i + \beta_2\texttt{Education}_i + \beta_3\texttt{ExpenditureYear0}_i + \beta_4\texttt{LabourForce}_i + \beta_5\texttt{StateSize}_i + \beta_6\texttt{Wage}_i + e_i$$

and the reduced model:

$$\log\left(\frac{\theta}{1-\theta}\right) = \beta_0 + \beta_1\texttt{Youth}_i + \beta_2\texttt{ExpenditureYear0}_i + e_i$$

(where θ is the probability that $\texttt{Y} > 100$) are shown in the Appendix. Perform a hypothesis test for whether the reduced model is a significantly poorer fit than the full model. Report a p-value and interpret the results. A few probability values are shown at the end of the Appendix. (8 points)

THE DIFFERENCE IN DEVIANCES FOR THE TWO MODELS IS $47.579 - 45.259 = 2.32$. AN APPROXIMATE P-VALUE CAN BE FOUND IN R BY $1 - \texttt{pchisq}(2.32, 4) = 0.6771$. BECAUSE THIS IS GREATER THAN $\alpha = 0.05$, WE CONCLUDE THAT THE REDUCED MODEL IS SUFFICIENT. THIS IS CONSISTENT WITH OUR FINDINGS WITH THE LINEAR REGRESSION MODEL ABOVE.

(b) Using the reduced model, compute the estimated probability that $\texttt{Y} > 100$ when `Youth` $= 130$ and `ExpenditureYear0` $= 60$. (8 points)

THE ESTIMATED LOG ODDS IS

$$-15.01848 + 0.06812 \times 130 + 0.06809 \times 60 = -2.0775$$

THUS, THE ESTIMATED PROBABILITY IS

$$\frac{\text{EXP}(-2.0775)}{1 + \text{EXP}(-2.0775)} = 0.1113$$

(c) Again using the reduced model, compute a 95% confidence interval for the odds ratio comparing two states with `Youth` = 130, one with `ExpenditureYear0` = 62 and the other with `ExpenditureYear0` = 60. (8 points)

THIS IS ASKING FOR A CONFIDENCE INTERVAL FOR $2\beta_2$. A 95% CONFIDENCE INTERVAL FOR THE LOG ODDS RATIO IS

$$2 \times 0.06809 \pm 1.96 \times 2 \times 0.02152 = [0.0518, 0.2205]$$

EXPONENTIATING THE ENDPOINTS GIVES A CONFIDENCE INTERVAL FOR THE ODDS RATIO: $[1.0532, 1.2467]$.

16. Consider a regression model with AR(1) correlated error terms:

$$y_t = \beta_0 + \beta_1 x_t + e_t$$

where $e_t = \rho e_{t-1} + \nu_t$ and $\nu_1, \nu_2, \ldots, \nu_n \overset{iid}{\sim} N(0, \sigma_\nu^2)$.

(a) Show that (4 points):

$$\text{Var}(e_t) = \sigma_e^2 = \frac{\sigma_\nu^2}{1 - \rho^2}$$

WE HAVE THAT

$$\text{VAR}(e_t) = \sigma_e^2 = \text{VAR}(\rho e_{t-1} + \nu_t) = \rho^2 \sigma_e^2 + \sigma_\nu^2$$

REARRANGING, WE THEREFORE HAVE

$$\sigma_e^2 = \frac{\sigma_\nu^2}{1 - \rho^2}$$

(b) Show that (5 points):

$$\text{Corr}(e_t, e_{t-1}) = \rho$$

WE HAVE THAT

$$e_t = \rho e_{t-1} + \nu_t$$

So

$$\text{Cov}(e_t, e_{t-1}) = \text{Cov}(\rho e_{t-1} + \nu_t, e_{t-1}) = \rho\text{Cov}(e_{t-1}, e_{t-1}) = \rho\sigma_e^2$$

which means that

$$\text{Corr}(e_t, e_{t-1}) = \frac{\text{Cov}(e_t, e_{t-1})}{\sigma_e^2} = \rho$$

(c) Consider fitting the above model using usual least squares (instead of, say, GLS) and obtaining an estimate of β_1, $\hat{\beta}_{1\text{LS}}$. Explain in words (no formulas required) whether and why confidence intervals for β_1 based on $\hat{\beta}_{1\text{LS}}$ (ignoring any autocorrelation) are valid. In your answer, you should mention the expected value and variance of $\hat{\beta}_{1\text{LS}}$. (8 points)

The expected value of $\hat{\beta}_{1\text{LS}}$ equals β_1. We saw in the notes that

$$\text{Var}\left(\hat{\beta}_{1\text{LS}}\right) = \frac{\sigma_i^2}{SXX}\left(1 + \frac{1}{SXX}\sum_{i \neq j}\sum \left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)\rho^{|i-j|}\right)$$

When $\rho = 0$, this equals $\frac{\sigma_e^2}{SXX}$, the usual variance of $\hat{\beta}_1$ for a least-squares model. Thus, the standard error of this statistic will be too small in the presence of positive correlation.

(d) Explain in words (no formulas required) how we can transform the above model so that the transformed model has *iid* errors. Why would we want to do this? (8 points)

We can use e.g. the Choleski decomposition to compute a "square root" of the Σ covariance matrix. Multiplying all terms in our model by the inverse of the square root matrix produces a model with *iid* errors. We do this in order to make residual plots meaningful. We can use least squares to fit the transformed model, then interpret residual plots for that model in the usual way to check for model validity.

Appendix

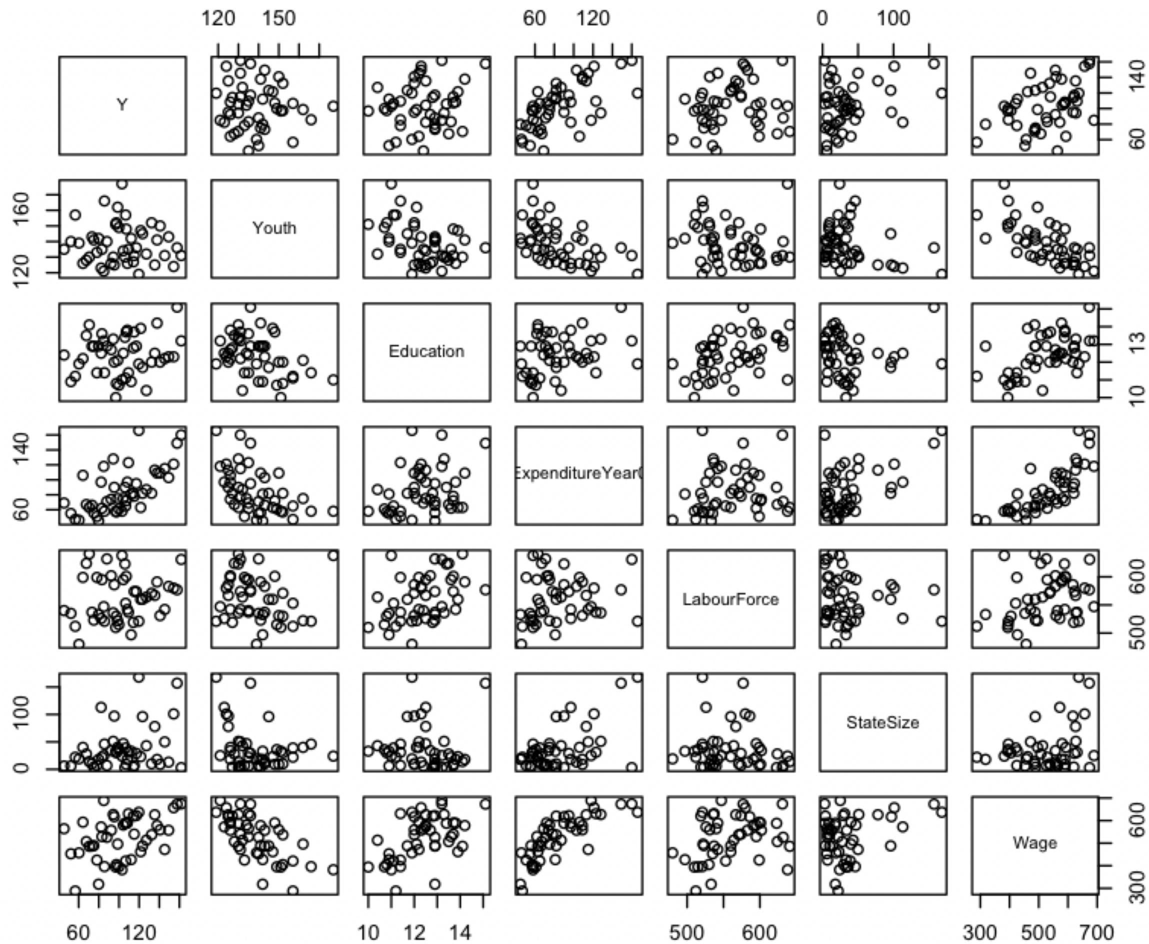


Figure 1: Pairs plot of crime data.

Sample correlation matrix of crime data:

	Y	Youth	Education	ExpenditureYear0	LabourForce	StateSize	Wage
Y	1.00	-0.06	0.16	0.65	0.17	0.31	0.42
Youth	-0.06	1.00	-0.40	-0.51	-0.16	-0.28	-0.67
Education	0.16	-0.40	1.00	0.30	0.43	0.00	0.52
ExpenditureYear0	0.65	-0.51	0.30	1.00	0.12	0.53	0.79
LabourForce	0.17	-0.16	0.43	0.12	1.00	-0.12	0.29
StateSize	0.31	-0.28	0.00	0.53	-0.12	1.00	0.31
Wage	0.42	-0.67	0.52	0.79	0.29	0.31	1.00

VIFs for crime data explanatory variables:

```
> vif(lm(Y ~ Youth + Education + ExpenditureYear0 + LabourForce + StateSize + Wage))
           Youth        Education ExpenditureYear0      LabourForce        StateSize             Wage 
        1.886849         1.610824         3.535904         1.287524         1.523856         4.477732 
```

Summary of fitted full and reduced linear regression models for crime data:

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -1.389e+02  8.008e+01  -1.735 0.090420 .  
Youth             8.650e-01  3.409e-01   2.537 0.015174 *  
Education         4.713e-01  3.533e+00   0.133 0.894555    
ExpenditureYear0  8.208e-01  1.974e-01   4.159 0.000164 ***
LabourForce       9.214e-02  8.758e-02   1.052 0.299067    
StateSize        -3.345e-03  1.011e-01  -0.033 0.973778    
Wage             -1.011e-02  6.840e-02  -0.148 0.883290    

Residual standard error: 21.15 on 40 degrees of freedom
Multiple R-squared:  0.5339,Adjusted R-squared:  0.4639 
F-statistic: 7.635 on 6 and 40 DF,  p-value: 1.678e-05

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -82.2122    44.7227  -1.838  0.07278 .  
Youth              0.8397     0.2793   3.007  0.00435 ** 
ExpenditureYear0   0.8078     0.1181   6.841 1.96e-08 ***

Residual standard error: 20.53 on 44 degrees of freedom
Multiple R-squared:  0.5169,Adjusted R-squared:  0.4949 
F-statistic: 23.54 on 2 and 44 DF,  p-value: 1.121e-07
```

The results of calling `regsubsets` on the full crime linear regression model:

```
         Youth Education ExpenditureYear0 LabourForce StateSize Wage
1  ( 1 ) " "   " "       "*"              " "         " "       " " 
2  ( 1 ) "*"   " "       "*"              " "         " "       " " 
3  ( 1 ) "*"   " "       "*"              "*"         " "       " " 
4  ( 1 ) "*"   " "       "*"              "*"         " "       "*" 
5  ( 1 ) "*"   "*"       "*"              "*"         " "       "*" 
6  ( 1 ) "*"   "*"       "*"              "*"         "*"       "*" 
```

p	2	3	4	5	6	7
BIC	297.46	292.53	294.73	298.57	302.40	306.25

Table 1: BIC values for the models with p explanatory variables.

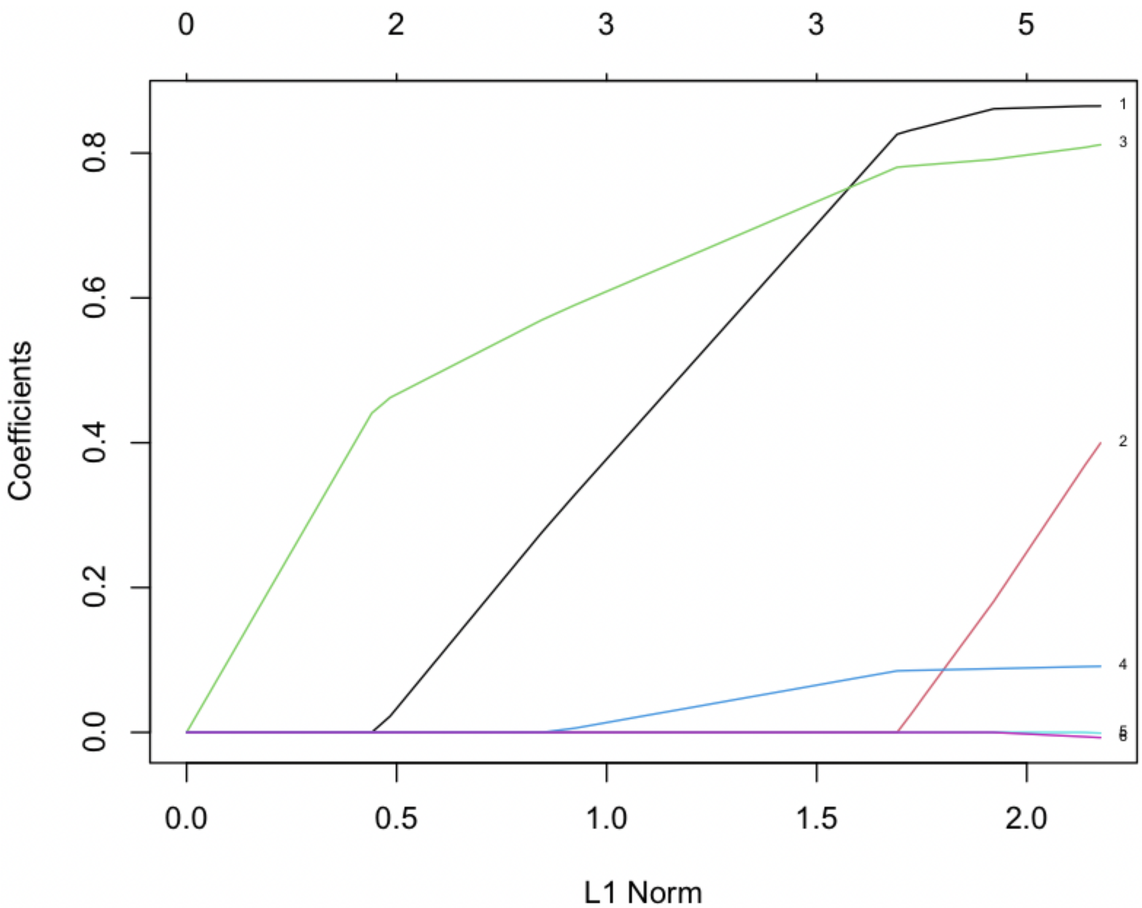


Figure 2: Plot of lasso regression coefficients.

Summary of fitted full and reduced logistic regression models for the crime data:

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -23.423753  10.921116  -2.145    0.032 *
Youth             0.081088   0.042944   1.888    0.059 .
Education        -0.028735   0.465414  -0.062    0.951
ExpenditureYear0  0.074010   0.031299   2.365    0.018 *
LabourForce       0.011025   0.011276   0.978    0.328
StateSize        -0.010574   0.014790  -0.715    0.475
Wage              0.001264   0.008698   0.145    0.884

Null deviance: 65.135  on 46  degrees of freedom
Residual deviance: 45.259  on 40  degrees of freedom
AIC: 59.259
```

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -15.01848    5.93289  -2.531  0.01136 *
Youth             0.06812    0.03390   2.009  0.04451 *
ExpenditureYear0  0.06809    0.02152   3.163  0.00156 **

Null deviance: 65.135  on 46  degrees of freedom
Residual deviance: 47.579  on 44  degrees of freedom
AIC: 53.579
```

A few probability values:

```
> 1 - pf(3.048100, 4, 40)
[1] 0.02772032
> 1 - pf(3.892841, 4, 40)
[1] 0.009204328
> 1 - pf(0.3645354, 4, 40)
[1] 0.8323904
> 1 - pf(0.218936, 4, 40)
[1] 0.926326
> 1 - pchisq(0.31, 4)
[1] 0.9891595
> 1 - pchisq(0.99, 4)
[1] 0.9113085>
> 1 - pchisq(2.08, 4)
[1] 0.7210476
1 - pchisq(2.32, 4)
[1] 0.6771302
```