

STATISTICS 608 Linear Models -EXAM I

April 1, 2014

Student's Name: _____

Student's Email Address: _____

INSTRUCTIONS FOR STUDENTS:

1. There are **14** pages including this cover page.
2. You have exactly 75 minutes to complete the exam.
3. There may be more than one correct answer; choose the best answer.
4. You will not be penalized for submitting too much detail in your answers, but you may be penalized for not providing enough detail.
5. Do not discuss or provide any information to any one concerning any of the questions on this exam or your solutions until I post the solutions next week.
6. You may use one 8.5" X 11" sheet of notes and a calculator.
7. At the end of the exam, leave your sheet of notes with your proctor along with the exam.

I attest that I spent no more than 75 minutes to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature: _____

INSTRUCTIONS FOR PROCTOR:

Immediately after the student completes the exam scan it to a pdf file and have student upload to Webassign.

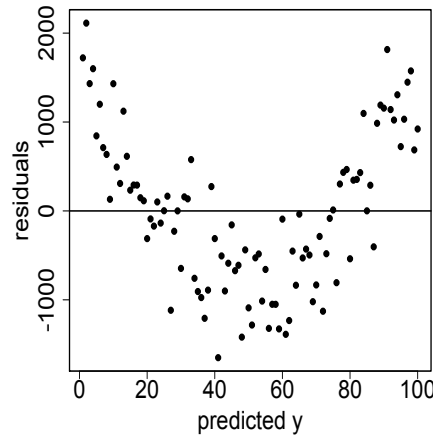
1. I certify that the time at which the student started the exam was _____ and the time at which the student completed the exam was _____.
2. I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
3. I certify that the exam was scanned in to a pdf and uploaded to Webassign in my presence.
4. I certify that the student has left the exam and sheet of notes with me, to be returned to the student no less than one week after the exam or shredded.

Proctor's Signature: _____

Part I: Multiple choice

1. An added variable plot is most useful for which of the following in a linear regression model?
 - (a) Visually assessing the effect of predictors after adjusting for the effects of the other predictors.
 - (b) Visually assessing whether the mean function is modeled appropriately.
 - (c) Determining whether the predictors are correlated with each other.
 - (d) Determining whether a polynomial term not currently included should be added to the model.
 - (e) Estimating an appropriate transformation of the predictors and/or response variable.
2. In a model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$, what does it mean for the predictor variable x_1 to have a variance inflation factor of 3?
 - (a) The variance of $\hat{\beta}_1$ is three times higher than it would have been if the predictors hadn't been linearly related.
 - (b) The largest standardized residual is 3, indicating we should investigate outliers in this model.
 - (c) The variance of the residuals is three times higher for some values of x_1 than for others.
 - (d) There is a "bad" outlier, a point with both a large standardized residual and high leverage, in this model that should be investigated.
3. A model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$ was fit to a data set, where x_1 was a continuous predictor and x_2 was an indicator or dummy variable. If all parameters are non-zero, what is the geometric interpretation of the model?
 - (a) A single regression line
 - (b) Parallel lines
 - (c) Two lines with the same y-intercept but separate slopes
 - (d) Two lines with separate intercepts and separate slopes
 - (e) Two parabolas with separate intercepts
4. In which of the following scenarios would we be most likely to attempt fitting a weighted least squares regression model with weights $w_i = n_i$?
 - (a) In a sample based on Y_i the average of n_i observations.
 - (b) In a model for which the residuals were not normally distributed.
 - (c) In a model for which a percent increase interpretation of slope was desired.
 - (d) In a data set for which variance inflation factors were high.

5. Suppose a multiple linear regression model has been fit, and the following plot of the residuals against the fitted values is created. Which of the following adjustments is suggested by the plot?



- (a) Transform some of the variables
 - (b) Include an interaction term
 - (c) Include polynomial terms for one or more predictors
 - (d) Remove outliers
 - (e) In general, it cannot be determined in what way the model is invalid based on residual plots.
6. A researcher has two models that are being compared:

$$\text{Model 1: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e, \text{ with } R^2 = 0.8$$

$$\text{Model 2: } y = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + e, \text{ with } R^2 = 0.7$$

The response variable is the same in both models, though the predictors used are different. The researcher states that because R^2 is larger in the first model, it is more valid than the second. Do you agree? Why or why not?

- (a) Yes, because the proportion of variability in y explained is greater for Model 1.
- (b) Yes, because Model 1 is more linear than Model 2.
- (c) Yes, because the higher value of R^2 indicates Model 1 has fewer outliers than Model 2.
- (d) No, because R^2 is not invariant under transformation.
- (e) No, because we need to check other plots like residual plots to find out whether model assumptions are met.

Part II: Long Answer

7. A study investigating distractions during lectures randomly assigned 100 students to each of the following groups (for a total of 300 students):

- Group 1 was permitted to use cell phones during lecture but not eat.
- Group 2 was permitted to eat but not use cell phones.
- Group 3 was not permitted to eat or use cell phones.

The response variable was the grade on a quiz following the lectures. The resulting design matrix used was:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where each entry was a 100×1 vector. The estimates of the three parameters, average grades on the quiz for each of the three groups, was $\hat{\boldsymbol{\beta}} = [42.81, 59.84, 58.67]$.

(a) Suppose researchers are interested in testing whether the average quiz grade for the first and third groups were equal. Write down the null and alternative hypotheses using vector notation, i.e. define \mathbf{a} for $\mathbf{a}'\boldsymbol{\beta}$.

(b) We saw in the homework that $\text{Var}(\mathbf{a}'\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}$. In our case, $\hat{\sigma}^2 = 105.06$

$$\text{and } (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.001 & 0 & 0 \\ 0 & 0.001 & 0 \\ 0 & 0 & 0.001 \end{bmatrix}.$$

Use this to calculate $\text{Var}(\mathbf{a}'\hat{\boldsymbol{\beta}}|\mathbf{X})$ for our problem.

8. Suppose that $\text{Var}(e_i|x_i) = \sqrt{n_i}\sigma^2$ for the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + e_i$ ($i = 1, 2, \dots, n$). Clearly define and write down a formula for a weighted estimate for the parameters β_0 and β_1 . There is no need to finish any algebra that you may set up; simply write down equations and define terms not already defined.

9. A clinical trial for hormone therapy was concerned in part with predicting the log of HDL cholesterol level using predictors age and body mass index (BMI). Researchers considered two models. For the first model, they noted BMI was right-skewed, and transformed it to normality using a log transformation. For the second, they considered adding a polynomial term in BMI. The two models are shown below; their marginal model plots are shown in the appendix. (Ignore the “c” in the “BMlc” on those plots for now.)

$$\text{Model 1: } \log(HDL) = \beta_0 + \beta_1 \text{ age} + \beta_2 \log(BMI) + e$$

$$\text{Model 2: } \log(HDL) = \beta_0 + \beta_1 \text{ age} + \beta_2 BMI + \beta_3 BMI^2 + e$$

- (a) If you had to choose one of the models based only on their marginal model plots, which would you prefer? Explain.
- (b) Researchers decided to use Model 2. Calculate a 95% confidence interval for β_1 from Model 2 and interpret your interval in context.

10. In a study modeling the progression of diabetes (*prog*), predictors age, BMI, sex, blood pressure (*bp*), and six serum measurements (*s1-s6*) were under consideration. The first model considered included all the predictors (output for Model 1 is included in the appendix):

$$\text{Model 1: } \textit{prog} = \beta_0 + \beta_1 \textit{age} + \beta_2 \textit{BMI} + \beta_3 \textit{sex} + \beta_4 \textit{bp} + \beta_5 s_1 + \beta_6 s_2 + \beta_7 s_3 + \beta_8 s_4 + \beta_9 s_5 + \beta_{10} s_6 + e$$

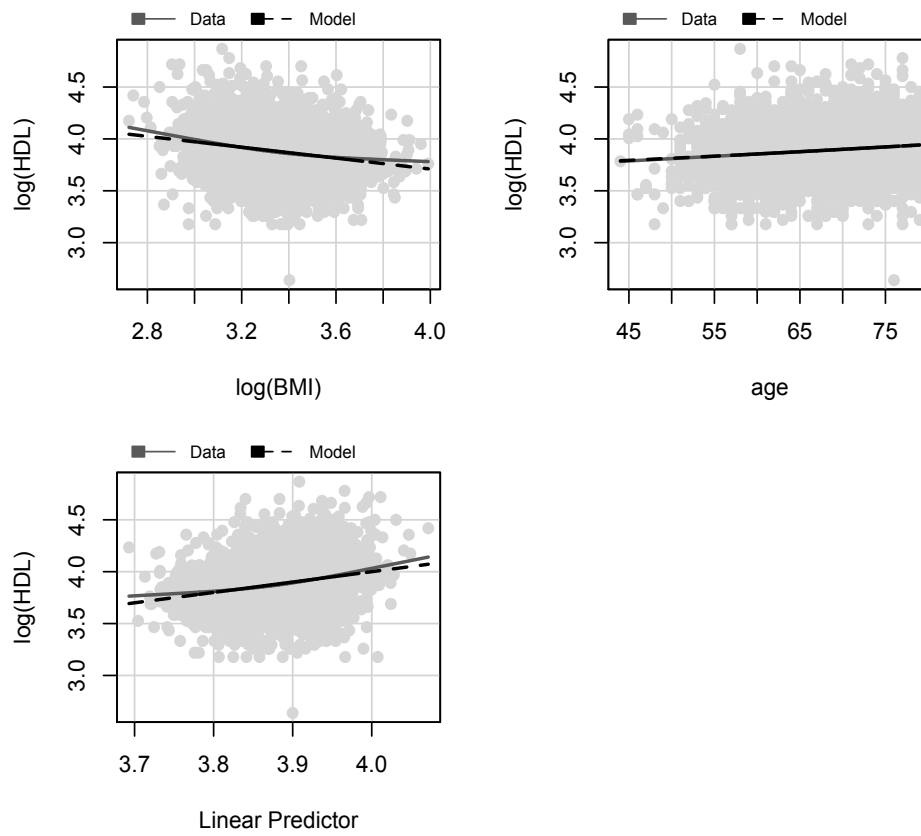
- (a) Is there any problem with multicollinearity for this model? Explain.
- (b) A researcher is interested in simply removing all variables in the model that have large p-values at one time. Is this method appropriate? Explain.

- (c) A second model under consideration is below. Conduct an F-test of model reduction to determine whether the model can be reduced, assuming appropriate model assumptions are met. (The p-value for the test is 0.1508.) Be sure to state your hypotheses, give a test statistic, and write a conclusion in the context of the problem.

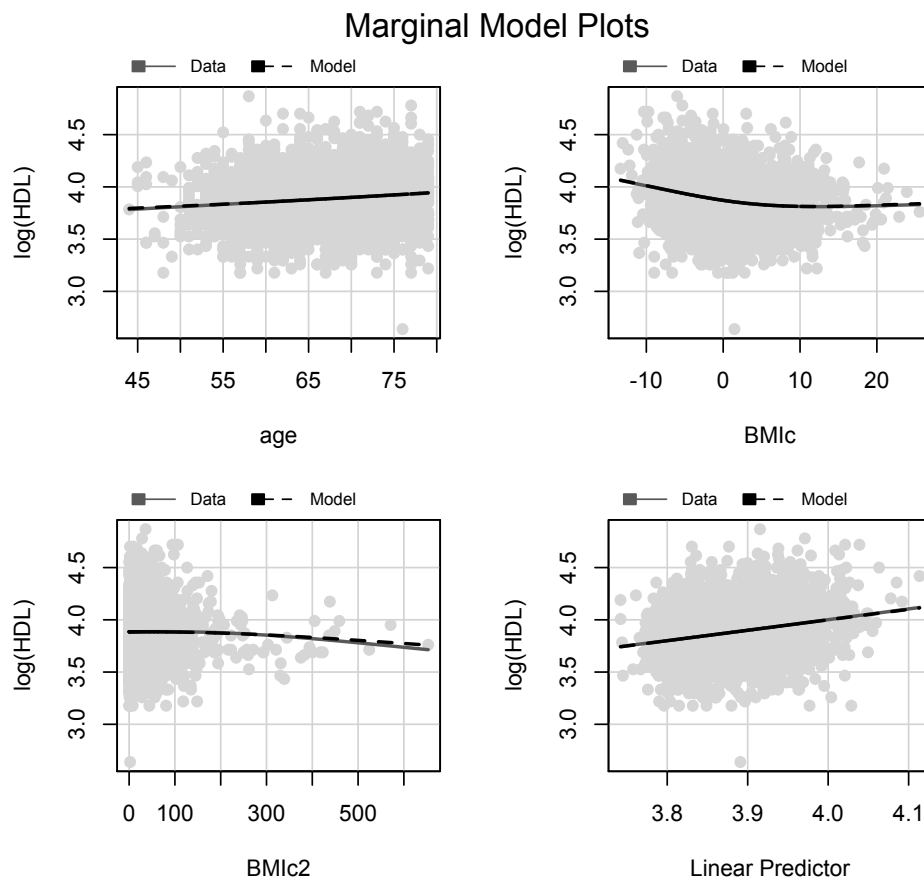
$$\text{Model 2: } \textit{prog} = \beta_0 + \beta_1 \textit{BMI} + \beta_2 \textit{sex} + \beta_3 \textit{bp} + \beta_4 \textit{s}_3 + \beta_5 \textit{s}_5 + e$$

HDL prediction: Model 1

Marginal Model Plots



HDL prediction: Model 2



HDL prediction: Model 2

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.6417749	0.0484669	75.139	< 2e-16 ***
age	0.0034526	0.0007219	4.783	1.82e-06 ***
BMic	-0.0096966	0.0009772	-9.923	< 2e-16 ***
BMic2	0.0004219	0.0001041	4.053	5.21e-05 ***

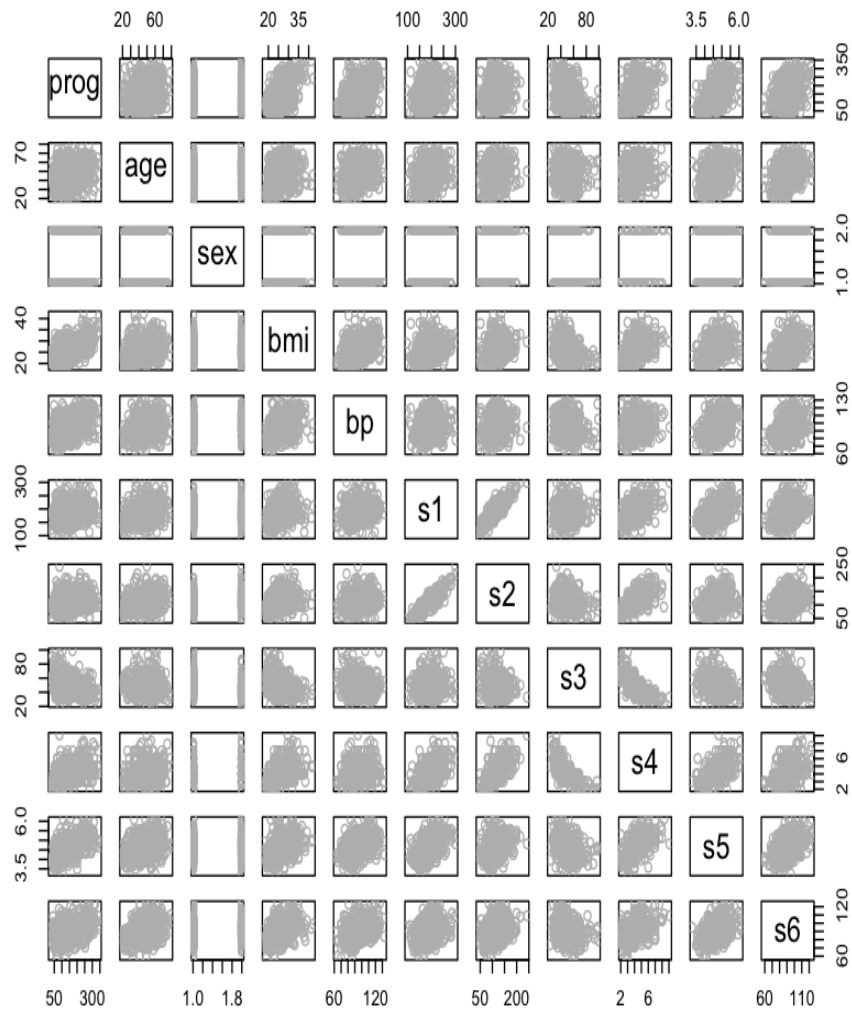
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.2482 on 2743 degrees of freedom

Multiple R-squared: 0.048, Adjusted R-squared: 0.04696

F-statistic: 46.1 on 3 and 2743 DF, p-value: < 2.2e-16

Diabetes Progression: Scatterplot Matrix



Diabetes Progression: Model 1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	VIF
(Intercept)	-334.56709	67.45462	-4.960	1.02e-06	
age	-0.03636	0.21704	-0.168	0.867030	1.217
sex	-22.85965	5.83582	-3.917	0.000104	1.278
bmi	5.60296	0.71711	7.813	4.30e-14	1.509
bp	1.11681	0.22524	4.958	1.02e-06	1.459
s1	-1.09000	0.57333	-1.901	0.057948	59.203
s2	0.74645	0.53083	1.406	0.160390	39.193
s3	0.37200	0.78246	0.475	0.634724	15.402
s4	6.53383	5.95864	1.097	0.273459	8.891
s5	68.48312	15.66972	4.370	1.56e-05	10.076
s6	0.28012	0.27331	1.025	0.305989	1.485

Residual standard error: 54.15 on 431 degrees of freedom

Multiple R-squared: 0.5177, Adjusted R-squared: 0.5066

F-statistic: 46.27 on 10 and 431 DF, p-value: < 2.2e-16

Analysis of Variance Table

Response: prog

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
age	1	92527	92527	31.5504	3.490e-08	***
sex	1	293	293	0.1000	0.7519	
bmi	1	826955	826955	281.9792	< 2.2e-16	***
bp	1	129312	129312	44.0934	9.448e-11	***
s1	1	1791	1791	0.6108	0.4349	
s2	1	5058	5058	1.7246	0.1898	
s3	1	237329	237329	80.9257	< 2.2e-16	***
s4	1	1821	1821	0.6210	0.4311	
s5	1	58856	58856	20.0690	9.582e-06	***
s6	1	3080	3080	1.0504	0.3060	
Residuals	431	1263986	2933			

Diabetes Progression: Model 2

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	VIF
(Intercept)	-217.6849	35.7638	-6.087	2.53e-09	
sex	-22.4742	5.7640	-3.899	0.000112	1.238
bmi	5.6431	0.7037	8.019	9.93e-15	1.443
bp	1.1232	0.2172	5.171	3.55e-07	1.347
s3	-1.0644	0.2417	-4.404	1.34e-05	1.459
s5	43.2344	5.9874	7.221	2.32e-12	1.461

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.35 on 436 degrees of freedom

Multiple R-squared: 0.5086, Adjusted R-squared: 0.503

F-statistic: 90.26 on 5 and 436 DF, p-value: < 2.2e-16

Analysis of Variance Table

Response: prog

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	1	4860	4860	1.6454	0.2003
bmi	1	896764	896764	303.5911	< 2.2e-16 ***
bp	1	146250	146250	49.5115	7.690e-12 ***
s3	1	131238	131238	44.4293	7.989e-11 ***
s5	1	154016	154016	52.1406	2.317e-12 ***
Residuals	436	1287881	2954		