Stat 740 Fall 2012: Midterm Exam, Tuesday October 16 Answer Key

There are four problems; write only on *one side of the paper* and start a new page for each problem.

1. The following is an ANOVA table from fitting a regression model is SAS with three continuous predictors x_1, x_2, and x_3 to a response Y; several parts of the table are missing. The sample size is $n = 19$.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	dfR	4966.67801	MSR	F	0.0021
Error	dfE	SSE	MSE		
Corrected Total	18	8087.68421			

   (a) What are the missing values of `dfR`, `dfE`, `SSE`, `MSR`, `MSE`, and `F` in the ANOVA table?
   Answer: dfR= $p - 1 = 4 - 1 = 3$, dfE= $n - p = 19 - 4 = 15$,
   SSE=SSTO−SSR= $8087.7 - 4966.7 = 3121$,
   MSR=$4966.7/3 = 1655.6$, MSE= $3121/15 = 208$,
   F=MSR/MSE= $1655.6/208 = 7.96$.

   (b) Compute and interpret R^2 for this model.
   Answer: R^2 =SSR/SSTO= $4966.7/8087.7 = 0.614$. 61.4% of the variability in the response is explained by the model.

   (c) If $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ is true, what does `MSR` estimate?
   Answer: σ^2.

   (d) What does the p-value for the F-test in the ANOVA table test? What do you conclude at the 5% level?
   Answer: The p-value tests $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$; we reject at the 5% level because p-value $= 0.0021 < 0.05 = \alpha$.

   (e) What is an estimate of σ?
   Answer: $\sqrt{\text{MSE}} = \sqrt{208} = 14.4$.

For the same data set, the following is the ANOVA table and table of regression effects from a fit of the model including *only the single predictor* x_2:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4519.89726	4519.89726	21.54	0.0002
Error	17	3567.78695	209.86982		
Corrected Total	18	8087.68421			

Parameter Estimates

Variable	DF	Estimate	Parameter Error	t Value	Standard Pr > |t|
Intercept	1	72.87601	7.19467	10.13	<.0001
x2	1	-0.67707	0.14590	-4.64	0.0002

(f) Compute $\text{SSR}(x_1, x_3|x_2)$. What does this statistic measure?
Answer: $\text{SSR}(x_1, x_3|x_2) = SSE(x_2) - SSE(x_1, x_2, x_3) = 3567.8 - 3121 = 446.8$. This is how much $\text{SSE}(x_2)$ is reduced by adding x_1 and x_3.

(g) Compute $R^2_{Y13|2}$. What does this statistic measure?
Answer: $R^2_{Y13|2} = SSR(x_1, x_3|x_2)/SSE(x_2) = 446.8/3567.8 = 0.125$. This is *the proportion* by which $\text{SSE}(x_2)$ is reduced by adding x_1 and x_3, not much here.

(h) What is the Pearson correlation between x_2 and Y?
Answer: For the reduced model, $R^2 = 4519.9/8087.7 = 0.559$, so $r = -\sqrt{0.559} \approx -0.75$. The negative sign comes from the sign of the regression coefficient -0.67707.

2. Short answer:

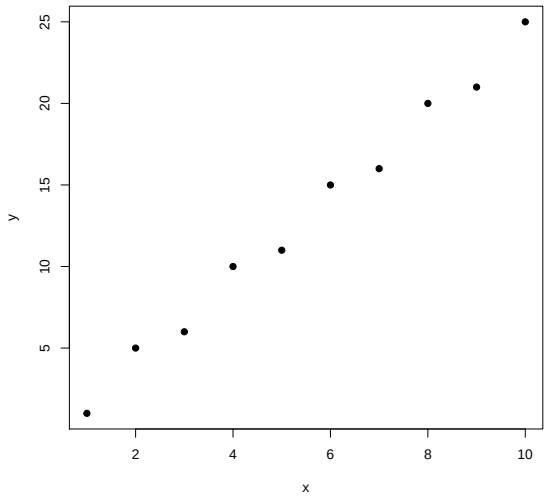
(a) Define the variance inflation factor VIF_2 for predictor x_2 in a model with three predictors x_1, x_2, x_3.
Answer: $\frac{1}{1-R^2}$, where R^2 is from regressing x_2 on x_1 and x_3.

(b) What assumptions does the Mann-Whitney-Wilcoxin test assume for testing $H_0 : F_1(x) = F_2(x)$?
Answer: Nothing.

(c) Where η is the population median, what assumptions does the signed rank test assume for testing $H_0 : \eta = 0$?
Answer: That the population density is symmetric about η.

(d) What assumptions does the sign test assume for testing $H_0 : \eta = 0$?
Answer: Nothing.

(e) What is the Spearman correlation of the data in the following scatterplot?



Answer: One; the data are monotone increasing.

(f) What can you say about the Pearson correlation for the data in part (e)?
Answer: It will be close to one, but a bit less than one.

(g) In what situation would you use the Spearman correlation instead of the Pearson correlation?
Answer: The Spearman correlation is robust to outlying observations, and measures a more general association, rather than just linear association.

(h) What type of relationships does a pairwise scatterplot matrix provide information on?
Answer: Marginal relationships.

(i) Let $\text{cov}(\mathbf{Y}) = \begin{bmatrix} 4 & -3 \\ -3 & 9 \end{bmatrix}$. Find $\rho_{12} = \text{corr}(Y_1, Y_2)$.
Answer:

$$\rho_{12} = \frac{-3}{\sqrt{4}\sqrt{9}} = -0.5.$$

(j) Define the t_ν random variable, i.e. the student's t with ν degrees of freedom.
Answer: Let $Z \sim N(0,1)$ independent of $X \sim \chi^2_\nu$. Then

$$T = \frac{Z}{\sqrt{X/\nu}} \sim t_\nu.$$

(k) Let $Y_i \overset{ind.}{\sim} N(\mu, \sigma_i^2)$ for $i = 1, 2, 3$ (mean is same, variance is different); here "*ind.*" means "independently." What is the distribution of $\bar{Y} = (Y_1 + Y_2 + Y_3)/3$?
Answer:

$$\bar{Y} \sim N\left(\mu, \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{9}\right).$$

3. Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ be the general linear model with $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \mathbf{I}_n\sigma^2)$. Let $\boldsymbol{\beta}$ be $p \times 1$.

   (a) What is the least squares estimator of $\boldsymbol{\beta}$, $\mathbf{b}$?
   Answer:
   $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

   (b) What is the distribution of $\mathbf{b}$?
   Answer:
   $\mathbf{b} \sim N_p(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$.

   (c) Show that $\mathbf{b}$ is unbiased, $E(\mathbf{b}) = \boldsymbol{\beta}$, directly from part (a) and $Y \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}_n\sigma^2)$.
   Answer:

$$\begin{aligned} E(\mathbf{b}) &= E\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\{\mathbf{Y}\} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} = \mathbf{I}_p\boldsymbol{\beta} = \boldsymbol{\beta}. \end{aligned}$$

   (d) How is the standard error of b_1 (the estimate of β_1) obtained? Assume $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$.
   Answer:
   $se(b_1) = \sqrt{v_2\ MSE}$ where v_2 is the $2nd$ diagonal entry in $(\mathbf{X}'\mathbf{X})^{-1}$.

4. In an experiment to determine the efficacy of a "quick acting" experimental steroid inhaler, one of two asthma medications was administered to $n = 138$ adult male volunteers. Each participant was assigned an asthma severity score x_{i2} on continuous scale from 0 (no asthma) to 20 (persistent, debilitating asthma), and an allergy severity index x_{i3} taking on values 1 (no allergies), 2 (mild, infrequent allergies), 3 (severe, frequent allergies), and 4 (constant, debilitating allergies). In the study, patients were cleared of asthmatic symptoms either by oral steroids or albuterol and given a standard preventative inhaler (treatment A, $x_{i1} = 0$), or the experimental inhaler (treatment B $x_{i1} = 1$). The patient then recorded the first instance of moderate to severe wheezing and the time to first wheezing was computed in hours Y_i.

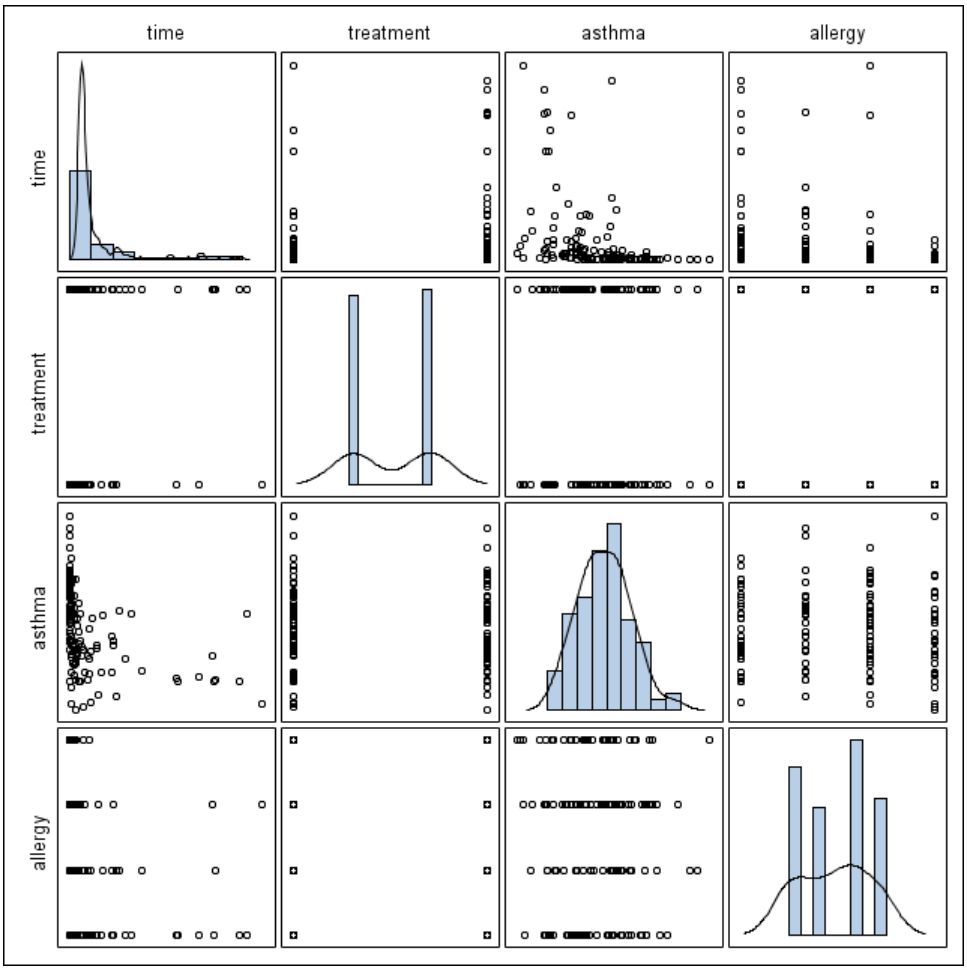
   You are to analyze these data keeping in mind the implied goal of the experiment. In particular, you are to determine if there is a treatment difference x_1 in the presence of the other concomitant variables x_2 and x_3. Treat allergy severity x_3 as *continuous*, not categorical. These data are on the web at `http://www.stat.sc.edu/~hansont/asthma.sas`.

   Start by examining marginal relationships via a scatterplot matrix, then fit the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

   Assess the fit of this model via residual plots and formally test constant variance using the Breusch-Pagan test. At this point, you may wish to consider one or more transformations, if necessary, to aid in your finding of an appropriate model. For your final model, discuss multicollinearity, model adequacy (i.e. model checking), and carefully interpret each regression coefficient. Clearly state your final model and

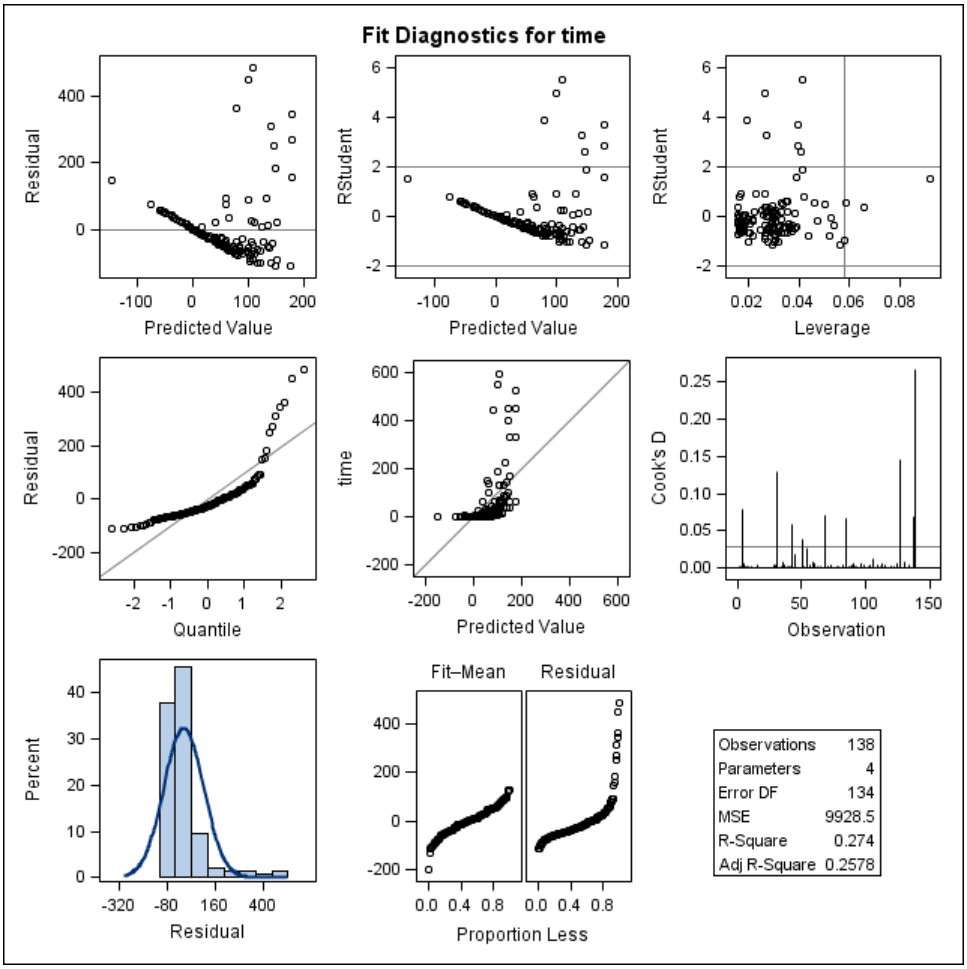
address the scientific question of whether the experimental inhaler increases time-to-wheezing. Provide and carefully interpret a 95% prediction interval for the time-to-wheezing in hours for a patient that receives the experimental inhaler (treatment B) with an allergy score of 3 and an asthma severity score of 15. Provide and interpret a 95% prediction interval for the time-to-wheezing in hours for a patient that receives the standard preventative inhaler (treatment A) treatment with an allergy score of 3 and an asthma severity score of 15. Contrast these two intervals.
Answer: The scatterplot matrix shows marginal nonlinearity and nonconstant variance of time versus asthma and allergy:



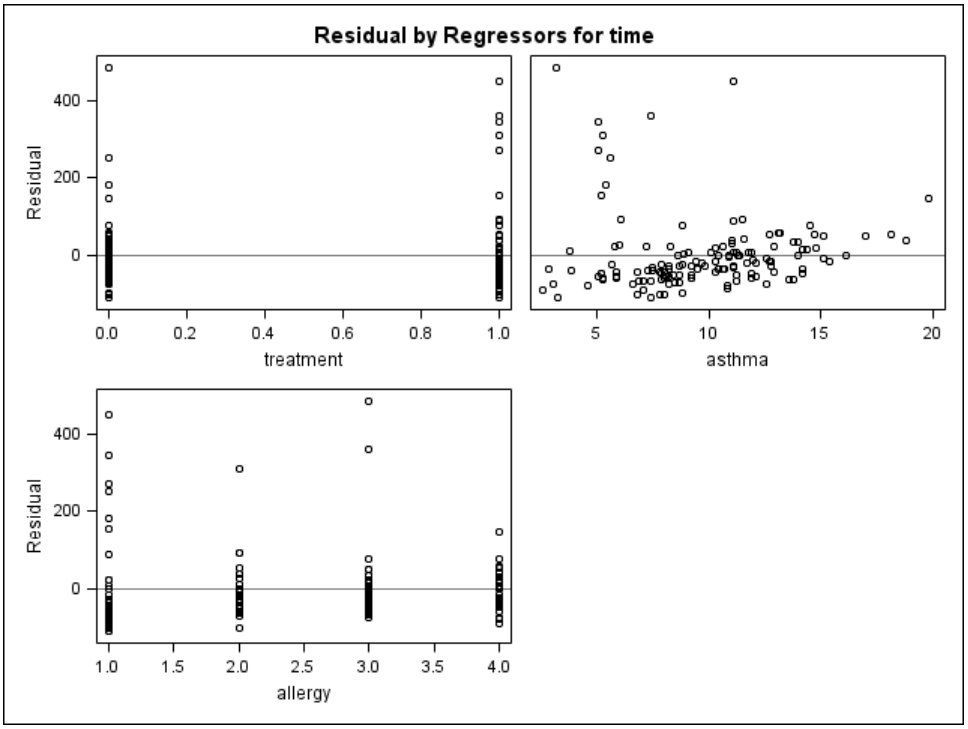
A fit of the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Shows nonconstant variance and nonlinearity in e_i vs. $\hat{Y}_i$, as well as rather severe non-normality from the Q-Q plot:



The e_i vs. x_{i2} also shows non-constant variance and a nonlinear trend:

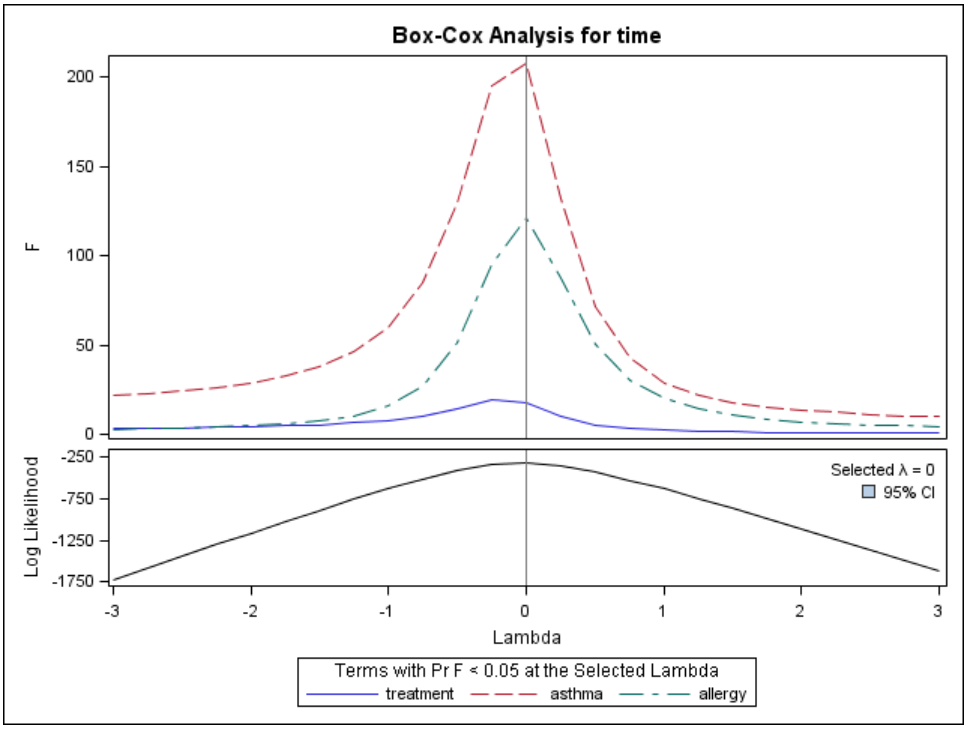


The Breusch-Pagan test in PROC MODEL gives

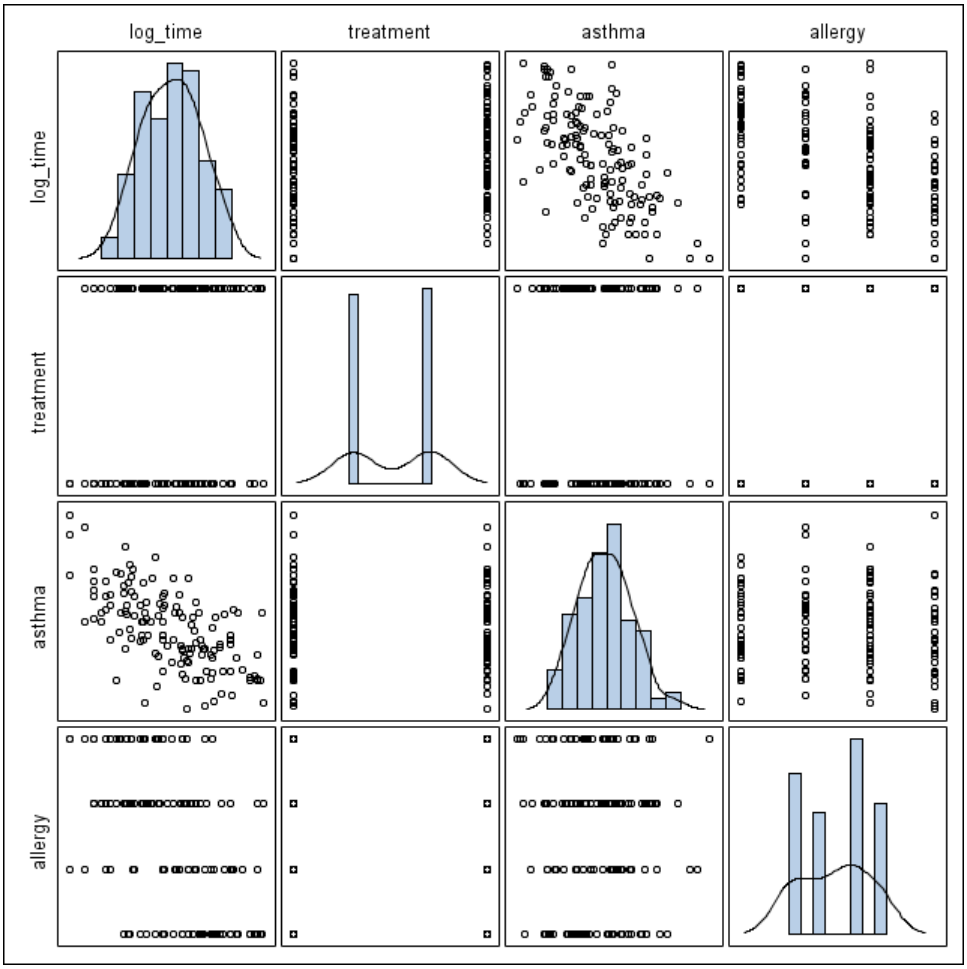
Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
time	Breusch-Pagan	11.22	3	0.0106	1, treatment, asthma, allergy

We reject $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ in the variance model $\sigma_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3}$ – there is significant evidence of nonconstant variance at the 5% level as $0.0106 < 0.05$.

The Box-Cox analysis gives



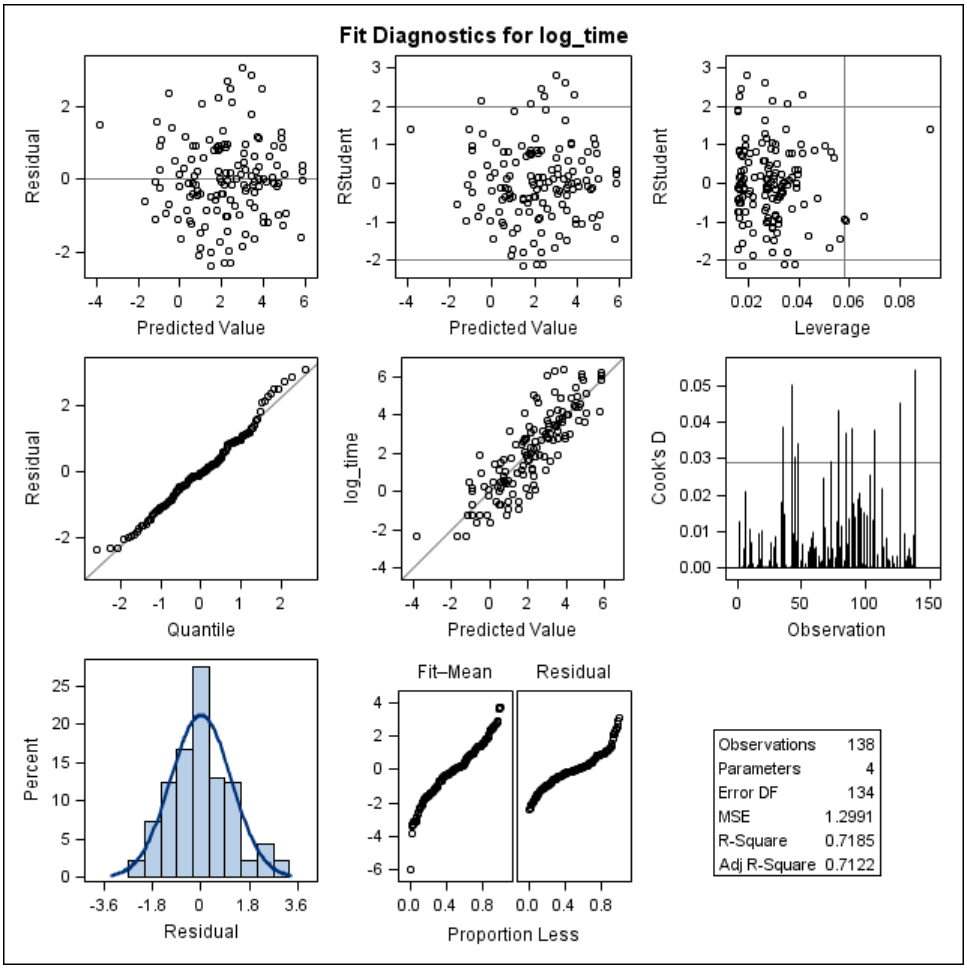
suggesting log(time) as the response. This transformation clears up nonconstant variance and shows roughly linear marginal relationships:



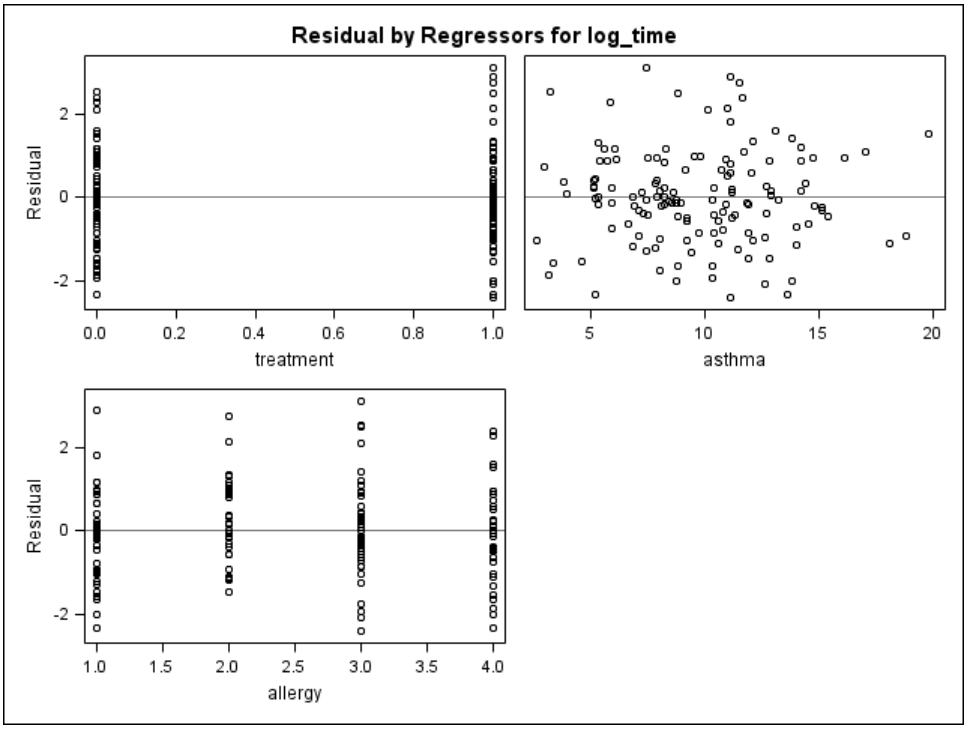
A fit of the model

$$\log(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Shows roughly constant variance and linearity in e_i vs. $\hat{Y}_i$, as well as reasonable normality of the e_i from the Q-Q plot:



The e_i vs. each of x_{i1}, x_{i2}, and x_{i3} shows roughly constant variance and no evidence of nonlinearity:



Dependent Variable: log_time

Number of Observations Read	140
Number of Observations Used	138
Number of Observations with Missing Values	2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	444.34159	148.11386	114.01	<.0001
Error	134	174.08391	1.29913		
Corrected Total	137	618.42550			

Root MSE	1.13980	R-Square	0.7185

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	8.09128	0.38748	20.88	<.0001	0
treatment	1	0.81772	0.19458	4.20	<.0001	1.00528
asthma	1	-0.40647	0.02820	-14.42	<.0001	1.00183
allergy	1	-0.96959	0.08813	-11.00	<.0001	1.00699

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
137	6.1150	4.8155	0.1879	2.5308	7.1003	1.2995
138	6.3881	3.8818	0.2318	1.5814	6.1823	2.5062
139	.	-0.9145	0.2082	-3.2061	1.3771	.
140	.	-0.0968	0.2113	-2.3895	2.1960	.

The experimental inhaler significantly ($p < .0001$ for testing $H_0 : \beta_1 = 0$) increases the time-to-wheezing by a factor of $\exp(0.8177) \approx 2.3$ times (i.e. 130%), holding asthma

and allergy severity constant. Increasing asthma severity one unit significantly *reduces* the time to wheezing by a factor of $\exp(-0.40647) \approx 0.67$ holding treatment and allergy constant. Increasing allergy severity one unit significantly *reduces* the time to wheezing by a factor of $\exp(-0.96959) \approx 0.38$ holding treatment and asthma severity constant. The three variables explain about $R^2 = 72\%$ of the variability in the log time-to-wheezing. The three variance inflation factors are all less than 10, if fact they are close to one, indicating there is no problem with multicollinearity.

I added the records

```
. 15 3 1
. 15 3 2
```

to the data step to get 95% prediction intervals for experimental treatment vs. control. For those on the standard inhaler with allergy=3 and asthma=15, there is 95% probability that their log(time) will be within the interval $(e^{-3.21}, e^{1.37}) = (0.04, 3.94)$ hours. For those on the experimental inhaler with allergy=3 and asthma=15, there is 95% probability that their log(time) will be within the interval $(e^{-2.3895}, e^{2.1960}) = (0.09, 8.99)$ hours. Note that the assumption of normality is crucial here, even with a large sample size, as we are considering prediction intervals.

```
options nocenter;

proc sgscatter;
 matrix time treatment asthma allergy / diagonal=(histogram kernel);
run;

proc model;
 parms beta0 beta1 beta2 beta3;
 time=beta0+treatment*beta1+asthma*beta2+allergy*beta3;
 fit time / breusch=(1 treatment asthma allergy);
run;

ods png; ods graphics on;
proc reg;
 model time=treatment asthma allergy;
run;
ods png; ods graphics on;

proc transreg;
 model boxcox(time) = identity(treatment asthma allergy);
run;

proc sgscatter;
 matrix log_time treatment asthma allergy / diagonal=(histogram kernel);
run;

ods png; ods graphics on;
proc reg;
 model log_time=treatment asthma allergy / cli vif;
run;
ods png; ods graphics on;
```