STAT 308 – In Class Exam 2

Name: Solutions

- There are a total of 50 points possible on this exam.
- The score of this in class exam will be combined with your take home exam for your total Exam 2 score.
- For questions where you are asked to perform calculations, you may leave your answers incomplete. For example $\frac{900-644}{7}$ would be an acceptable answer.
- You are allowed one two-sided "cheat sheet" for use on this exam.
- If you fill out the course evaluations at the end of the semester, you will receive 1 bonus point on this exam.

Good luck!

This page intentionally left blank.

1. [2 pts each] **TRUE OR FALSE** For each of the following statements, determine if they are true or false (please write out either true or false, so your answer is clear to me). If the statement is false, either explain why it is false or correct the statement to make it true.

a. For every variable I add to a regression model, the r^2 will always increase.

True

b. An estimate of a regression model's variance, σ^2, is the same as the mean squared error of the model.

True

c. Suppose we perform a test for significance of an overall regression model with 6 explanatory variables, and we reject the null hypothesis that the model is not significant. This means that we have statistically significant evidence that all six explanatory variables are significant in our regression model.

False, at least one explanatory variable is significant.

d. Looking at a residual plot, if the spread of the residuals gets larger as the fitted values get larger, we may want to consider a log transformation on the response variable.

True

e. If I use AIC as a variable selection criteria, we want to choose the regression model that maximizes the AIC.

False, we want to minimize AIC.

2. [3 pts each] Consider the following information on the degrees of freedom and sums of squares from an ANOVA table for a multiple linear regression model with 1 categorical explanatory variable, 2 numeric explanatory variables, and no polynomial terms or interactions.

	df	Sum Sq
X_1	3	112.33
X_2	1	33.22
X_3	1	55.28
Error	24	158.96

a. How many possible values are there for the categorical variable X_1?

$$4 = 3+1$$

b. Calculate an F statistic for testing for significance of an overall linear model with X_1, X_2, and X_3.

$$F = \frac{MSM}{MSE} = \frac{(112.33+33.22+55.28)/5}{158.96/24}$$

c. What are the numerator and denominator degrees of freedom for the F statistic in (b)?

5 and 24

d. What is the mean squared error for the regression model with just X_1 as an explanatory variable?

$$MSE = \frac{33.22+55.28+158.96}{24+1+1}$$

e. Calculate an F statistic for testing if adding X_3 to a model that already includes X_1 and X_2 significantly improves the predictive ability of the response?

$$F = \frac{MSM}{MSE} = \frac{55.28/1}{158.96/24}$$

3. An agricultural company wants to determine if different types of insecticides perform differently at repelling insects from crops. The observed dataset provides information on 72 experimental units of land that were sprayed with one of 6 different types of insecticide (labelled A-F), and the insect count on each experimental unit was recorded.

The R output needed for this problem is found at the back of this exam

a. [1 pt] What insecticide does R use as the baseline?

Spray A

b. [2 pts] What is the expected insect count for units sprayed with insecticide A?

14.5 insects

c. [2 pts] What is the expected insect count for units sprayed with insecticide D?

14.5 - 9.5833 insects

d. [2 pts] What is the difference in expected insect count for units sprayed with insecticide B vs. insecticide A?

0.833 insects

e. [2 pts] What is the difference in expected insect count for units sprayed with insecticide C vs. insecticide B?

-12.4167 - 0.833 insects

f. [4 pts] Perform a formal test to determine if the expected insect count is different for the different types of insecticides. Be sure to include all pieces of information needed to perform a hypothesis test. Assume $\alpha = 0.05$.

$H_0: \beta_1 = \beta_2 = \cdots = \beta_5 = 0$

H_A: At least one $\beta \neq 0$

$F \simeq 34.7$ $p = 0$ We reject H_0 and can conclude that a linear model with spray is statistically significant at predicting insect count.

4. A research laboratory is inteterested in predicting the survival time for leukemia patients based on their white blood cell (wbc) count at the time of diagnosis as well as the presence or absence of a morphologic characteristic of white blood cells (ag). The research fits a linear model for survival time with wbc and ag as the explanatory variables.

a. [2 pts] Consider the residual plot and QQ plot for the linear model in the R output. Are the assumptions for homoscedasticity and normally distributed residuals met? Why or why not?

The assumption of normality appears to be met because the points fall close to the 45° line. The assumption of homoscedasticity appears to be violated because the points appear to spread as fitted values get larger.

b. To combat these potential issues, I decide instead to fit a log-linear model. That is I transform the survival time to a log scale and fit that transformed response to a linear model with wbc and ag as the explanatory variables. Use the output under the log-linear model to answer these questions.

- i. [2 pts] Interpret the value of the parameter agpresent in the context of the problem.

  Holding white blood cell count constant, the expected log-survival time for leukemia patients is 1.0939 log months longer for patients where the characteristic is present.

- ii. [2 pts] Interpret the value of the parameter wbc in the context of the problem.

  When white blood cell count increases by 1, the expected log survival time decreases by 1.715×10^{-5} log months

- iii. [4 pts] Perform a formal hypothesis test to determine if the log-linear model for survival time with white blood cell count and the AG characteristic is significant. Be sure to include all pieces of information needed to perform a hypothesis test. Assume $\alpha = 0.05$.

  $H_0: \beta_1 = \beta_2 = 0$ H_A: At least one $\beta \neq 0$

  $F = 5.72$ $p = 0.0072$

  We reject H_0 and say that a log-linear model for survival time with wbc and the morphologic characteristic is significant.

- iv. [2 pts] Consider the residual plot and QQ plot for the log-linear model in the R output. Are the assumptions for homoscedasticity and normally distributed residuals met? Why or why not?

  Yes, the residual plot appears to be evenly spread across 0 and the QQ-plot follows close to the 45° line

Linear Model Output from the Insecticide Dataset

Table of the summary output

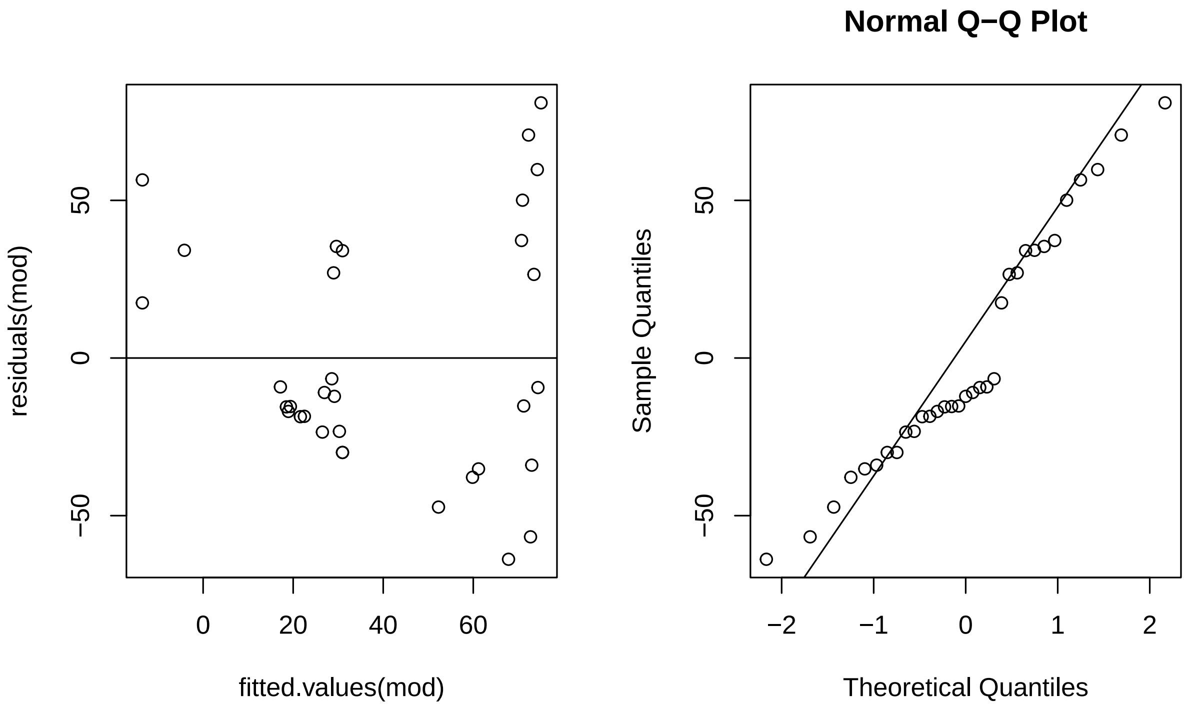
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	14.5000	1.1322	12.81	0.0000
sprayB	0.8333	1.6011	0.52	0.6045
sprayC	-12.4167	1.6011	-7.76	0.0000
sprayD	-9.5833	1.6011	-5.99	0.0000
sprayE	-11.0000	1.6011	-6.87	0.0000
sprayF	2.1667	1.6011	1.35	0.1806

ANOVA table for differences between the two models

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	71	3684.00				
2	66	1015.17	5	2668.83	34.70	0.0000

Linear Model Output from the Leukemia Dataset

Plots for Checking Assumptions



Log-linear Model Output from the Leukemia Dataset

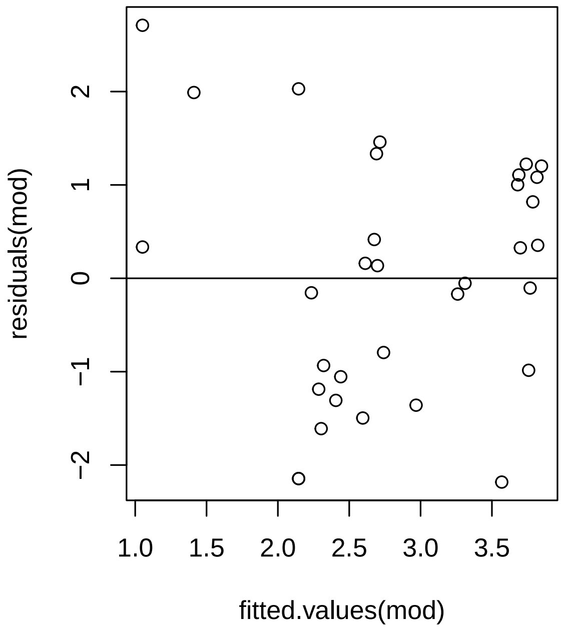
Table of the summary output

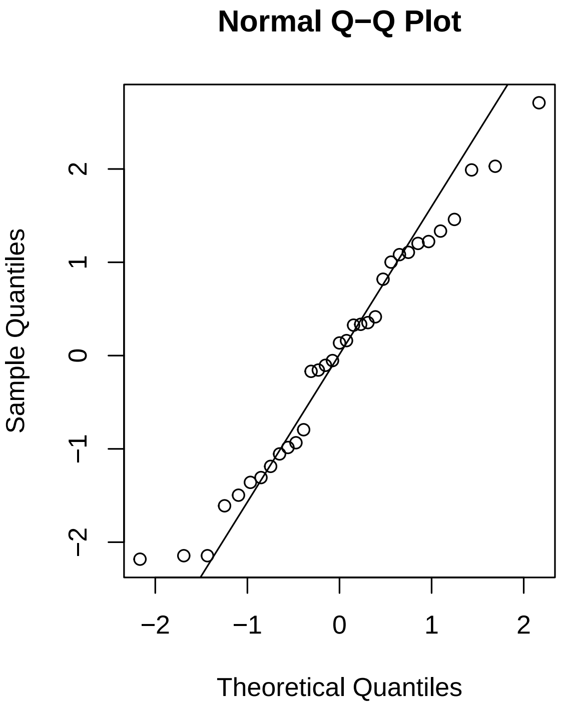
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.7665	0.3923	7.05	0.0000
agpresent	1.0939	0.4687	2.33	0.0265
wbc	-1.715e-5	6.897e-6	-2.49	0.0187

ANOVA table for differences between the two models

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	32	75.44				
2	30	54.32	2	21.12	5.83	0.0072

Plots for Checking Assumptions


residuals(mod)
fitted.values(mod)


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles