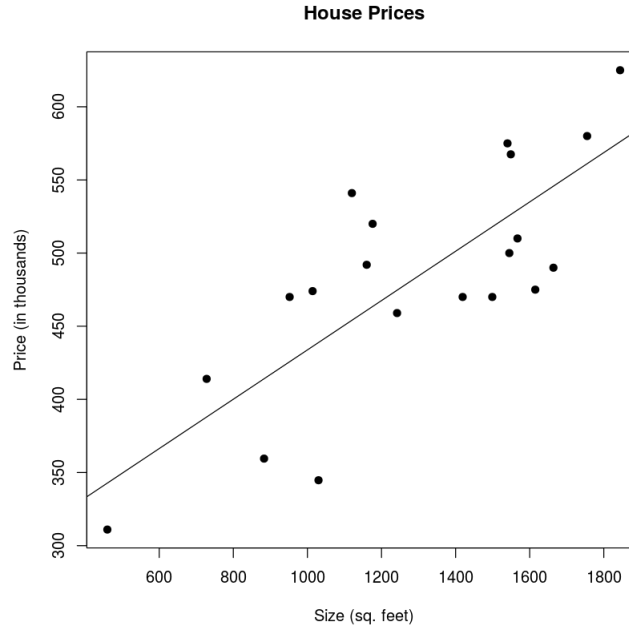


Statistical Methods - Math 222

Midterm 4 Review Problems

- In each case, state the specific statistical procedure that is appropriate for the given situation. Be specific: identify the response variable and the explanatory variable(s). If there are any categorical variables present, state how many levels each categorical variable has.
 - You want to study whether men and women get different average amounts of sleep at night.
 - You want to predict life satisfaction (on a scale of 1 to 10) based on several factors, including income, regional cost of living, commuting time, and number of children.
 - You want to predict whether a potential new store will succeed (make a profit) in a town based on several factors, including population of the town, average household income, and overall home-ownership rate.
- The scatterplot below shows the relationship between size (in square feet) and price (in thousands of dollars) of a random sample of 20 houses sold recently in Arroyo Grande, CA.



Below is a summary of the least squares regression model for this scatterplot.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	265.22212	42.64202	6.220	7.21e-06 ***
myData\$Size	0.16859	0.03188	5.288	5.00e-05 ***

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

Residual standard error: 51.31 on 18 degrees of freedom

Multiple R-squared: 0.6084, Adjusted R-squared: 0.5866

- Is the trend statistically significant? How can you tell?

- (b) If $SE_{\hat{\mu}} = 55.18$, find a 95% confidence interval for the mean home price of a 1200 square foot house.
- (c) Find a 95% prediction interval for the price of a 1200 square foot house (recall that $SE_{\hat{y}}^2 = SE_{\hat{\mu}}^2 + s^2$ where s is the residual standard error).
3. This example is based on data from 78 seventh-grade students in a rural midwestern school. The researcher was interested in the relationship between the students' "self-concept" and their academic performance. The data included each student's grade point average (GPA), score on a standard IQ test, and gender, taken from school records. Gender is coded as 1 for female and 2 for male. The final variable is each student's score on the Piers-Harris Children's Self-Concept Scale, a psychological test administered by the researcher. Below is a summary of the multiple linear regression model for this data in R.

Call:

```
lm(formula = gpa ~ iq + gender + concept, data = myData)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-3.5769 -0.7493  0.1984  0.9577  2.4089
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.83463     1.28584   -2.205 0.030641 *
iq           0.08079     0.01336    6.045 5.78e-08 ***
gender      -0.82214     0.31354   -2.622 0.010630 *
concept      0.05048     0.01396    3.616 0.000548 ***
---

```

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

Residual standard error: 1.323 on 73 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.561, Adjusted R-squared: 0.543

F-statistic: 31.1 on 3 and 73 DF, p-value: 4.643e-13

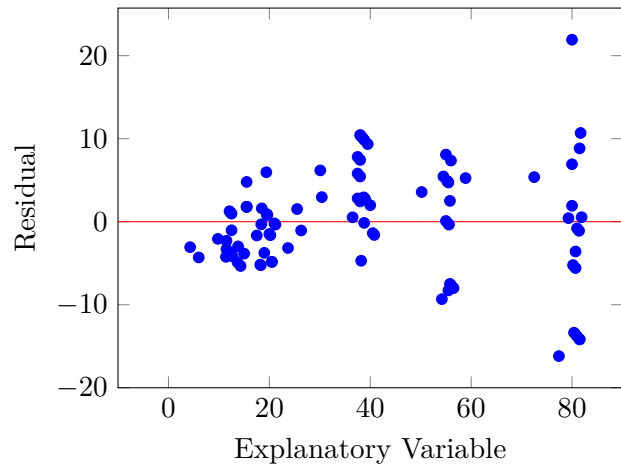
- (a) What is the formula for predicting GPA from IQ, Gender, and Self-Concept using this regression model?
- (b) What percent of the variability in GPA is explained by this model?
- (c) Describe in words the affect of each of the explanatory variables on the response variable.
4. A sample of 180 introductory statistics students found a correlation of $R = 0.61$ between student scores on midterm 1 versus midterm 2. The scores on midterm 1 had a standard deviation $s_x = 14.52$, while the scores on midterm 2 had a standard deviation of $s_y = 13.81$. A linear regression model for predicting the midterm 2 grade based on midterm 1 is:

$$y = 29.66 + 0.58x.$$

- (a) Use the information above to fill in the values in the ANOVA table:

	Degrees of Freedom	Sum of Squares	Mean Square	F
Model				
Residuals				n/a
Total				n/a

- (b) What is the residual standard error for this model?
5. Which two assumptions of linear regression can be most easily evaluated using a residual plot? Evaluate those assumptions for this plot.



6. To study factors that affect the recurrence of heart attacks (HA), an investigator collected data from 20 HA victims. The investigator fit a logistic regression model with an indicator of a second HA within one year (1 = HA; 0 = no HA) as the binary outcome. There are two predictors:
- $x_1 = 1$ if the patient completed an anger management program; 0 otherwise
 - $x_2 = \text{anxiety score}$ (0 = low anxiety, 100=high anxiety)

Computer output is given below:

	Estimate	Std. Error	z value	Pr(> z)
Intercept	-6.36347	3.21362	-1.980	0.0477
x1	-1.02411	1.17101	-0.875	0.3818
x2	0.11904	0.05497	2.165	0.0304

- (a) In terms of x_1 and x_2 , what are the odds of a patient having a second heart attack?
- (b) What is the probability of a second heart attack for a patient that has completed an anger management program and scored a 100 on the anxiety test?
- (c) Is there statistical evidence that an anger management program is associated with a reduction in the probability of a second heart attack? Explain.