Department of Mathematics

The University of Toledo

Master of Science Degree
Comprehensive Examination

Applied Statistics

April 23, 2016

Instructions

Do all four problems.

Show all of your computations.
Prove all of your assertions or quote appropriate theorems.
Books, notes, and calculators may be used.
This is a three hour test.

MS Applied Exam

Linear Regression Model

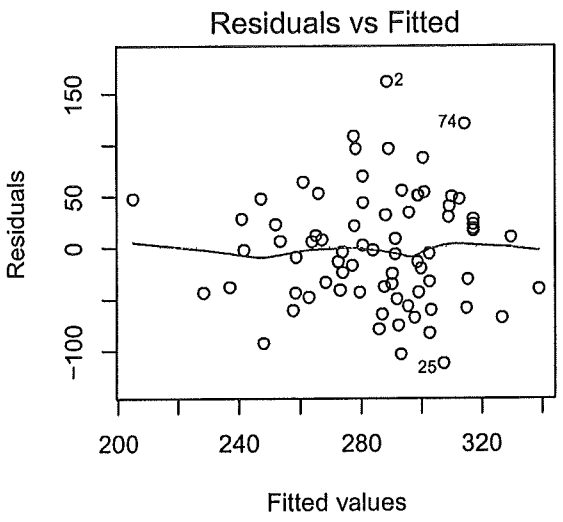
1. We are interested predicting the selling prices of homes in south Eugene, Oregon. We will predict selling price in thousands of dollars (`Price`, y) using floor space in thousands of square feet (`Floor`, x_1), number of bathrooms (`Bath`, x_2), number of bedrooms (`Bath`, x_3), and age in years since built (`AGE`, x_4)

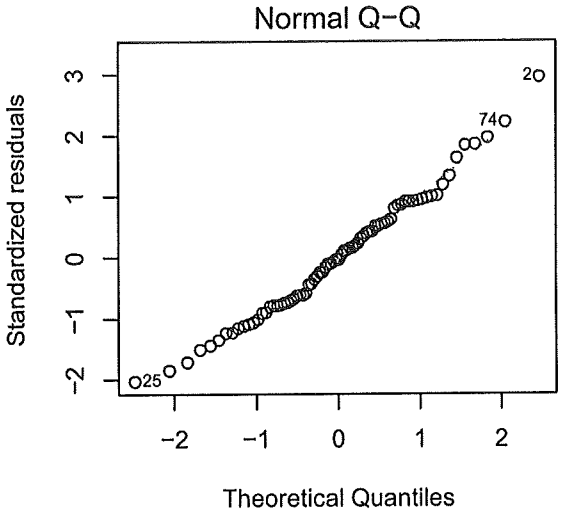
$$\text{Model 1: } y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i, \epsilon_i \sim^{iid} N(0, \sigma^2)$$

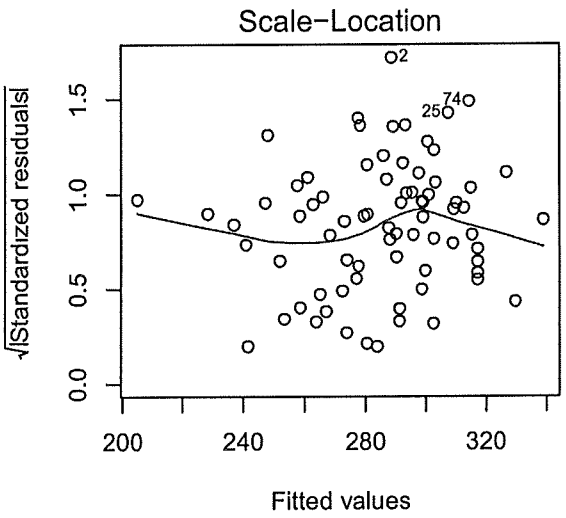
The model is fitted using the software R and the following summary output and diagnostic plots are obtained:

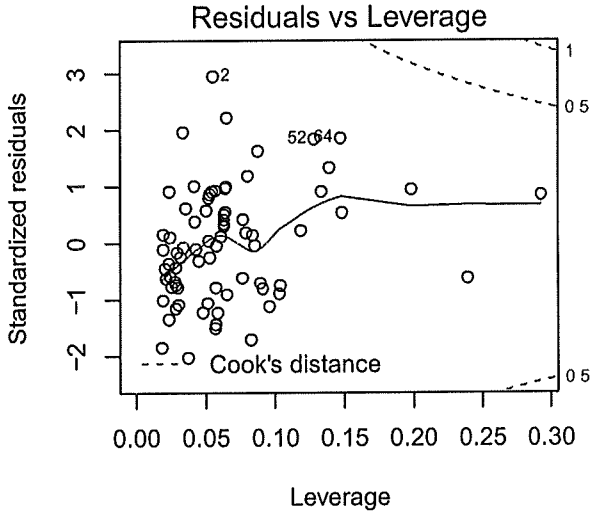
********** Model 1: y on x_1, x_2, x_3, x_4

```
Call:
lm(formula = Price ~ Floor + Bath + Bed + Year)

Residuals:
     Min       1Q   Median       3Q      Max
-112.405  -41.434   -2.113   35.716  161.550

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   ???      650.3838   0.975  0.33308
Floor         68.3907   35.6961   ???    0.05940 .
Bath          15.4014   13.7673   1.119  0.26704
Bed          -32.2435   10.2115  -3.158  0.00234 **
Year          -0.2060    0.3302  -0.624  0.53471
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 56.34 on 71 degrees of freedom
Multiple R-squared:  0.1744, Adjusted R-squared: ???
F-statistic: 3.749 on 4 and 71 DF,  p-value: 0.007991
```









(1) (2 pts) What is the value of the t-statistic of $\hat{\beta}_1$?

(2) (2 pts) How many observations are there in the data set?

(3) (2 pts) Has the null hypothesis $H_0 : \beta_3 = 0$ to be rejected on a 5% level? Why?

(4) (6 pts) How to interpret $\hat{\beta}_3$? Please write a sentence using the value and the variable names. What's the 95% confidence interval for β_3? How to interpret the confidence interval? If you believe you don't have enough information to calcualte the specific number, please write down the expression for calculation and denote the quantity(or quantities) you need.

(5) (4 pts) What is the estimate of the intercept $\hat{\beta}_0$? What is the adjusted R square?

(6) (2 pts) What is the estimate of $Var(\epsilon_i)$?

(7) (5 pts) Can you find the sample variance for response y, i.e. $Var(y)$? (*Hint*: you might need to use $R^2 = 1 - RSS/TSS$)

(8) (5 pts) What is $Var(y|x_1, x_2, x_3, x_4)$? Is this the same as $Var(y)$? Why or why not?

(9) (5 pts) What are the assumptions for linear regression model? Take a look at the residual plots. Are the model assumptions on the ϵ_i fullfilled? If not, what is the main problem?

(10) (4 pts) You want to repeat the regression, but with a better model and/or adapted data basis. What actions will you consider to take?

On the next few pages you are given the results from fitting three other models. Each model is labeled by a number and on the same line (with asterisks) the model is described.

(11) (10 pts) For each hypothesis below indicate whether or not the hypothesis can be tested from the model output given. Look at all **four** of the models fit before you answer (Model 1 at the begining, Model 2, 3 and 4 at the end of this question). For each hypothesis that **can** be tested, carry out the test (or just write the test statistic and use appropriate notation(s) if you think some information is missing), and tell me your conclusion. If you think the hypothesis cannot be tested, please state the reason. All β's here refer specifically to **Model 1**.

- i. $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

- ii. $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

- iii. $H_0 : \beta_2 = \beta_3 = 0$

- iv. $H_0 : \beta_1 = 0$

- v. $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$

********** Model 2: y on x_1

```
##
## Call:
## lm(formula = Price ~ Floor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.547  -40.849   -7.604   48.484  159.420
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   173.06      64.09   2.700  0.00858 **
## Floor          57.22      32.34   1.769  0.08099 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.49 on 74 degrees of freedom
## Multiple R-squared:  0.04058,    Adjusted R-squared:  0.02761
## F-statistic:  3.13 on 1 and 74 DF,  p-value: 0.08099
```

********** Model 3: y on x_1, x_4

```
##
## Call:
## lm(formula = Price ~ Floor + Year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -129.322  -40.397   -5.789   40.069  163.834
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -433.8578   576.7642  -0.752    0.454
## Floor         51.0780    32.8301   1.556    0.124
## Year           0.3143     0.2969   1.059    0.293
##
## Residual standard error: 59.45 on 73 degrees of freedom
## Multiple R-squared:  0.05509,    Adjusted R-squared:  0.0292
## F-statistic: 2.128 on 2 and 73 DF,  p-value: 0.1264
```

********** Model 4: y on x_2, x_3, x_4

```
##
## Call:
## lm(formula = Price ~ Bath + Bed + Year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -106.473  -38.291   -7.003   33.357  157.756
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 577.9510   661.6698   0.873  0.38531
## Bath         25.5236    12.9468   1.971  0.05252 .
## Bed         -27.7970    10.1270  -2.745  0.00764 **
## Year         -0.1283     0.3337  -0.385  0.70174
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57.38 on 72 degrees of freedom
## Multiple R-squared:  0.1317, Adjusted R-squared:  0.09551
## F-statistic:  3.64 on 3 and 72 DF,  p-value: 0.01666
```

2. Suppose random vector $\mathbf{X} = (X_1, X_2)' \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

(1) (3 pts) Suppose $\boldsymbol{\mu} = (7, 11)'$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 9 & -2 \\ -2 & 2 \end{pmatrix}$. Find the joint distribution of $Z_1 = X_1 + X_2$ and $Z_2 = X_1 - X_2$. Are Z_1 and Z_2 independent? Explain your reasoning.

(2) (3 pts) Suppose $\mathbf{X}_1, \ldots, \mathbf{X}_4 \sim^{ind} N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ have the same value as in part (1). Find the joint distribution of $V_1 = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 + \mathbf{X}_4$ and $V_2 = \mathbf{X}_1 - \mathbf{X}_4$. Are V_1 and V_2 independent? Explain your reasoning.

(3) In pratice, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are usually unknown. A random sample of size $n = 30$ yields the following mean and covariance matrix:

$$\bar{\mathbf{x}} = \begin{pmatrix} 6 \\ 12 \end{pmatrix} \text{ and } \mathbf{S} = \begin{pmatrix} 20/3 & -2/3 \\ -2/3 & 2/2 \end{pmatrix}$$

Hint: The inverse of matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is $\frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

Note: You may consult a table of t, or F distribution at the end, as appropriate. If you think you don't have enough information to find the confidence interval or perform hypothesis testing, denote what you need and answer the question with appropriate notation(s).

(a) (3 pts) Determine the sample mean and variance of $X_1 + X_2$ and $X_1 - X_2$. Also find the sample covariance of these two variables.

(b) (4 pts) At the significance level $\alpha = 0.05$, test

$$H_0 : \boldsymbol{\mu} = (7, 11)' \text{ vs } H_a : \boldsymbol{\mu} \neq (7, 11)'$$

(c) (3 pts) Give a 95% confidence region for $\boldsymbol{\mu}$. Is $\boldsymbol{\mu} = (8, 12)'$ in the confidence region above?

(d) (8 pts) Determine the 90% simultaneous T^2 confidence intervals for $\mu_1, \mu_2, \mu_1 + \mu_2$ and $\mu_1 - \mu_2$.

(e) (8 pts) Determine the 90% simultaneous Bonferroni confidence intervals for $\mu_1, \mu_2, \mu_1 + \mu_2$ and $\mu_1 - \mu_2$.

Critical Values you might need: You can tear this page for your convenience.

1. The following is the selected upper tail quantiles of t distribution, with selected degree of freedom (shown on the column names) and selected upper tail probability (shown in the row names)

n=	25	26	27	28	29	30
0.0125	2.385	2.379	2.373	2.368	2.364	2.360
0.025	2.060	2.056	2.052	2.048	2.045	2.042
0.05	1.708	1.706	1.703	1.701	1.699	1.697
0.075	1.485	1.483	1.482	1.480	1.479	1.477
0.01	1.316	1.315	1.314	1.313	1.311	1.310

2. The followings are the selected upper tail quantitles of the F distribution, for selected numbers of degrees of freedom. Columns give the first d.f, rows the second.

0.0125:

	1	2	3	4
24	7.291	5.289	4.469	4.009
25	7.244	5.248	4.431	3.972
26	7.202	5.211	4.396	3.938
27	7.162	5.177	4.363	3.907
28	7.126	5.145	4.334	3.879
29	7.093	5.116	4.307	3.852
30	7.062	5.089	4.281	3.828

0.025:

	1	2	3	4
24	5.717	4.319	3.721	3.379
25	5.686	4.291	3.694	3.353
26	5.659	4.265	3.670	3.329
27	5.633	4.242	3.647	3.307
28	5.610	4.221	3.626	3.286
29	5.588	4.201	3.607	3.267
30	5.568	4.182	3.589	3.250

0.05:

	1	2	3	4
24	4.260	3.403	3.009	2.776
25	4.242	3.385	2.991	2.759
26	4.225	3.369	2.975	2.743
27	4.210	3.354	2.960	2.728
28	4.196	3.340	2.947	2.714
29	4.183	3.328	2.934	2.701
30	4.171	3.316	2.922	2.690

0.075:

	1	2	3	4
24	3.464	2.891	2.607	2.434
25	3.452	2.878	2.594	2.421
26	3.440	2.866	2.582	2.408
27	3.430	2.855	2.571	2.397
28	3.420	2.845	2.560	2.387
29	3.411	2.836	2.551	2.377
30	3.402	2.827	2.542	2.368

0.1:

	1	2	3	4
24	2.927	2.538	2.327	2.195
25	2.918	2.528	2.317	2.184
26	2.909	2.519	2.307	2.174
27	2.901	2.511	2.299	2.165
28	2.894	2.503	2.291	2.157
29	2.887	2.495	2.283	2.149
30	2.881	2.489	2.276	2.142

3. (20 pts. Total, 2 pts. Each part) Here is a 2x4 contingency table with a chi-square analysis for the variables gender and BMI (Body Mass Index) for n = 93 members of an introductory statistics class in the 1970's. BMI is divided into four classes by its quartiles so that subjects would be evenly spread among the four BMI groups. The goal is to use this data to determine if there is a difference in the distribution of BMI between males and females, and if so, how are they different? Regard the subjects as one random sample of all students at that university for that year, where we measured two variables – gender and BMI (from height and weight). Note the *four quantities shown in italics* for each of the eight cells.

Rows: Gender Columns: BMI Quartile Group

	Q1	Q2	Q3	Q4
Male	8	12	18	20
	13.79	20.69	31.03	34.48
	14.97	13.72	14.97	14.34
	-6.968	-1.720	3.032	5.656
Female	16	X1	6	3
	45.71	X2	17.14	8.57
	9.03	X3	9.03	8.66
	6.968	1.720	-3.032	X4

Cell Contents:
Count
% of Row
Expected count
Residual

Pearson Chi-Square = 16.750, DF = X5, P-Value = X9

a. We could have used BMI cutoffs defined by the CDC (Centers for Disease Control) that define underweight, normal, and overweight. Out of the 93 subjects, 6 were classified as underweight, 80 as normal, and 7 as overweight. Give at least one reason favoring the use of quartile groups and at least one reason favoring the use of the CDC cutoffs.
b. X1 is the count for Females in the second quartile BMI group (between Q1 and the median). Find X1. Recall that the overall n was given.
c. Calculate X2.
d. State the null hypothesis for the chi-square test for this problem. Define parameters as needed to make this clear. Is this a test of independence (of the two variables) or of homogeneity (of the distribution of BMI for males & females)? Hint: I believe that both answers can be defended. Choose your favorite and defend it.
e. X3 is the expected count under the null hypothesis. Calculate X3. Hint: there are three ways to do it. Do two of them to check your answer.
f. X4 is the (raw) residual for females in the upper quarter of their BMI data. Calculate X4.
g. Note that the chi-square statistic has been calculated for you, but what is its degrees of freedom, df = X5? Show, and explain, the calculation. This is a function of the number of rows and the number of columns. Show that function. But ...
h. The df is also related to the total number of cells X6, the number of cells whose values are determined by the values in other cells X7, and the number of parameters requiring that they be estimated under the null hypothesis X8. Give the formula for df as a function of X6, X7 & X8, and find df in this problem. Hopefully it matches the answer in part g. Hint: these answers actually depend upon your answer to part d.
i. Find the (approximate) P-value X9. Interpret the result; draw your conclusion. What assumptions were necessary for this to be valid?
j. Assuming that you rejected the null hypothesis, how you can answer the second question stated in the overall goal – how do the distributions differ? Describe the difference. Hint: use the residuals (maybe a plot?).

4. (30 pts. total) One might think that the above problem could (or should) be answered by looking at the raw BMI data rather than dividing the subjects into BMI groups. This problem explores this possibility. Once again the goal is to learn whether or not the two genders differ in terms of the distribution of BMI, and then to explain how they differ if they do. Here are some basic statistics for BMI broken down by gender.

Variable	Gender	N	Mean	StDev	CoefVar	Minimum	Q1	Median	Q3	Maximum
BMI	Male	58	22.252	2.062	9.26	18.653	20.520	22.099	23.649	29.159
	Female	35	20.361	2.014	9.89	16.726	19.369	20.176	21.434	26.453

a. (6 pts.) Use the information given to create side-by-side boxplots. Identify outliers using "fences" at 1.5 times the IQR (Inter-quartile Range) beyond the quartiles. Are there any outliers?

b. (3 pts.) What is required for the two-sample t-test to be valid? Based on the information you have, do you believe that these assumptions are satisfied here? Why or why not?

c. (1 pt.) If we wish to use the easy formula where we pool the variances, what else is required?

d. (5 pts.) Test for equal variances between the two groups. Use level of significance $\alpha = .10$. Define parameters, state the hypotheses, and find the test statistic, the critical value and the P-value. What is your conclusion?

e. (5 pts.) Whatever you believe, or have found, about assumptions, do the pooled variance two-sample t-test to find out if we have good evidence that the distributions of BMI differ for males and females. Once again, define parameters, state the hypotheses, and find the test statistic, the critical value and the P-value. What is your conclusion?

f. (3 pts.) Find a 95% confidence interval for the difference in the mean BMI, Male – Female.

g. (7 pts.) For the sake of comparison, here is the computer output for the Mann-Whitney analysis. W is the Mann-Whitney Test Statistic, equal to the sum of the ranks of Group 1 (Males) in the pooled data. What is the parameter of interest (ETA)? What assumptions are required for this to be valid? What are the expected mean and standard deviation of the statistic under the null hypothesis (assume no ties)? What is the (approximate) null distribution of W and therefore what is the P-value? What is your conclusion? How do these results compare with the "parametric" analysis that you performed above? Are they similar, or more than a little different? Which analysis should we prefer, and why?

```
Point estimate for ETA1-ETA2 is 1.887
95.0 Percent CI for ETA1-ETA2 is (1.035,2.728)
W = 3239.0
```