Stat 5000 Homework #10
Fall 2024 due Fri, November 22nd @ 11:59 pm Name: Sam Olson
Collaborators: **The Hatman**

Q1

In the last homework assignment, the relationship between weight and height of 18 yearold girls was examined. For this assignment, you will examine additional variables collected in the Berkeley Guidance Study. The data are posted in the file BGSgirls2.txt. There is one line of data for each of 70 girls with the variables appearing in the following order:

- ID: Girl identification number
- WT2: Weight (kg) at 2 years
- HT2: Height (cm) at 2 years
- WT9: Weight (kg) at 9 years
- HT9: Height (cm) at 9 years
- LG9: Leg circumference (cm) at 9 years
- ST9: Strength (kg) at 9 years
- WT18: Weight (kg) at 18 years
- HT18: Height (cm) at 18 years
- LG18: Leg circumference (cm) at 18 years
- ST18: Strength (kg) at 18 years
- BMI: Body Mass Index at 18 years
- SOMA: Somatotype (SOMA), on a scale from 1, very thin, to 7, very obese

The goal of this exercise is to determine how well the measurements at ages 2 and 9 can predict BMI at age 18.

Use R to complete the following exercises:

(a)

Compute the sample correlations between BMI at age 18 and each of the following explanatory variables HT2, HT9, WT2, WT9, and ST9. Which of these explanatory variables have significant correlations with BMI at age 18?

(b)

Compute the sample correlations among the five explanatory variables HT2, HT9, WT2, WT9, and ST9. Which of these explanatory variables have significant correlations with other explanatory variables?

(c)

Find least squares estimates of the parameters in the regression of BMI at age 18 on strength at age 9,

$$BMI_i = \beta_0 + \beta_1 ST9_i + \epsilon_i, \text{ for } i = 1, ..., 70$$

Is the slope significantly different from zero? What do the residual plots reveal?

(d)

Now compute the multiple regression of the body mass index at age 18 on both weight at age 9 and strength at age 9, i.e. fit the model

$$BMI_i = \beta_0 + \beta_1 ST9_i + \beta_2 WT9_i + \epsilon_i, \text{ for } i = 1, ..., 70$$

Is the estimate of β_1 for this model, the coefficient for strength at age 9, the same as the estimate of β_1 for the model in part (c)? Did you expect the estimates to be different? Explain. Is the effect of strength at age 9 significant in this model?

(e)

Fit the multiple regression model

$$BMI_i = \beta_0 + \beta_1 WT2_i + \beta_2 HT2_i + \beta_3 WT9_i + \beta_4 HT9_i + \beta_5 ST9_i + \epsilon_i, \text{ for } i = 1, ..., 70$$

Report and interpret in context the R^2 value.

(f)

Report estimates of the six partial regression coefficients for the model in part (e), their standard errors, and the value of the corresponding t-tests and p-values (for two-sided alternatives to the null hypothesis). For each t-test, explicitly state the null hypothesis that is tested and interpret the result in context.

Q2

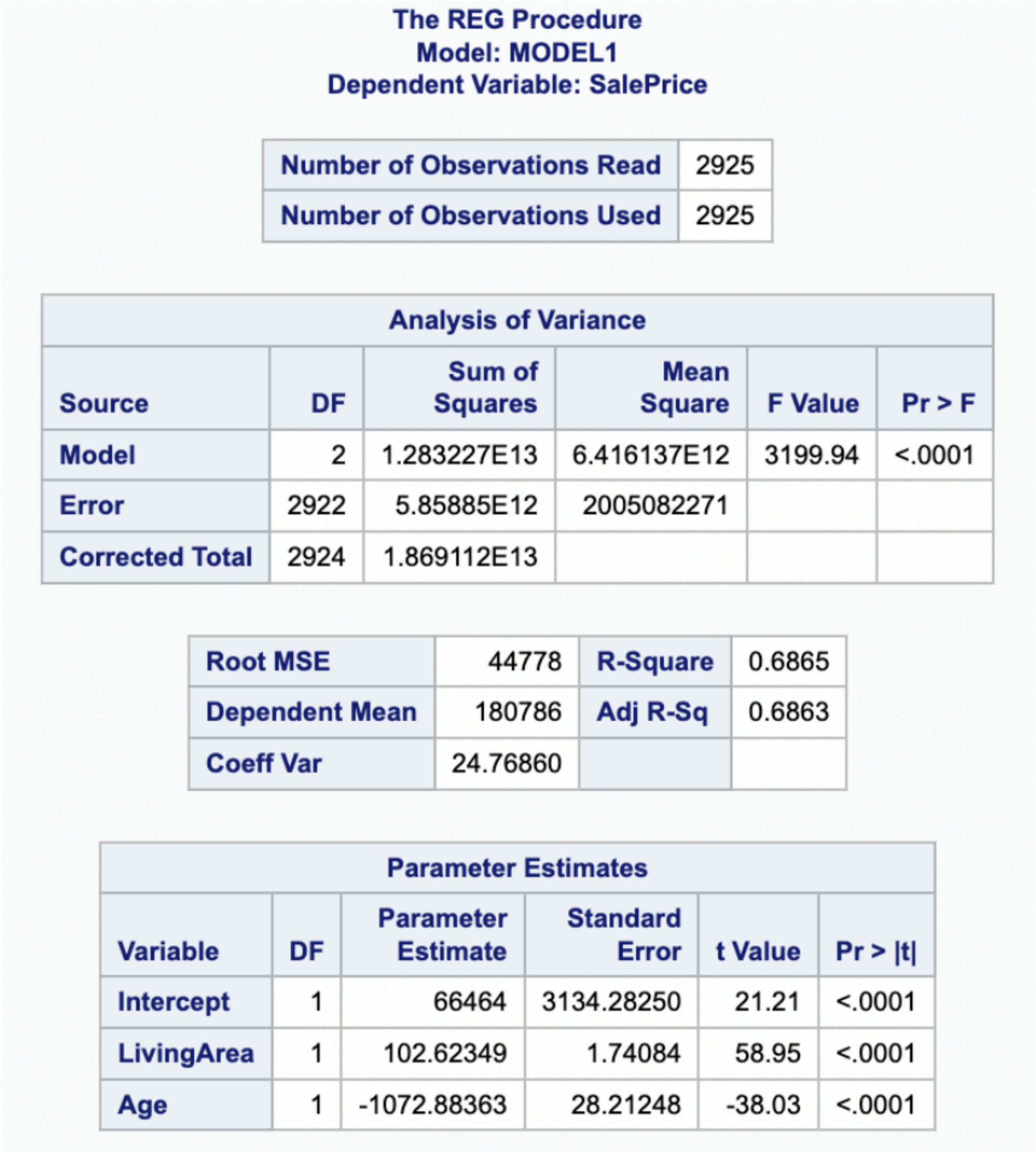
A dataset was collected from home sales in Ames, Iowa between 2006 and 2010. The variables collected are:

- Year Built: The year the house was built
- Basement Area (in sq. ft): The amount of area in the house below ground level
- Living Area (in sq. ft): The living area in the home (includes Basement Area)
- Total Room: The number of rooms in the house
- Garage Cars: The number of cars that can be placed in the garage
- Year Sold: The year the home was sold
- Sale Price: The sale price of the home (the response variable)
- Garage Size: S = Small (Garage Cars = 0,1) or L = Large (Garage Cars = 2+)
- Age (in yrs.): Age of house = Year Sold - Year Built

The data from 2,925 sales can be found in the file AmesHousing.csv posted in our course's shared folder on SAS Studio. For all parts requiring a hypothesis test, make sure to state the null and alternative hypotheses, test statistic, p-value, decision, and conclusion in context.

First, we will use SAS to explore predicting the Sale Price of a house from two explanatory variables: Living Area and Age. The SAS code that generated the output below is included in Canvas in the housing solution.sas file for your reference. Use the output shown on the next page to complete the exercises that follow.

(a)



The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2925
Number of Observations Used	2925

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.283227E13	6.416137E12	3199.94	<.0001
Error	2922	5.85885E12	2005082271		
Corrected Total	2924	1.869112E13			

Root MSE	44778	R-Square	0.6865
Dependent Mean	180786	Adj R-Sq	0.6863
Coeff Var	24.76860		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	66464	3134.28250	21.21	<.0001
LivingArea	1	102.62349	1.74084	58.95	<.0001
Age	1	-1072.88363	28.21248	-38.03	<.0001

Figure 1: CocoMelon

Give a description of the parameters (β's) for Living Area and Age in the multiple linear regression model.

(b)

What is the value of R^2 and its interpretation for the model including Living Area and Age?

(c)

Using the ANOVA table, conduct the F-test for the overall significance of the model. Report the null and alternative hypotheses, test statistic and p-value, and interpret the result in context.

(d)

Give the t-test for the significance of each explanatory variable in the model. Report the null and alternative hypotheses, test statistic and p-value, and interpret the result in context.

i.

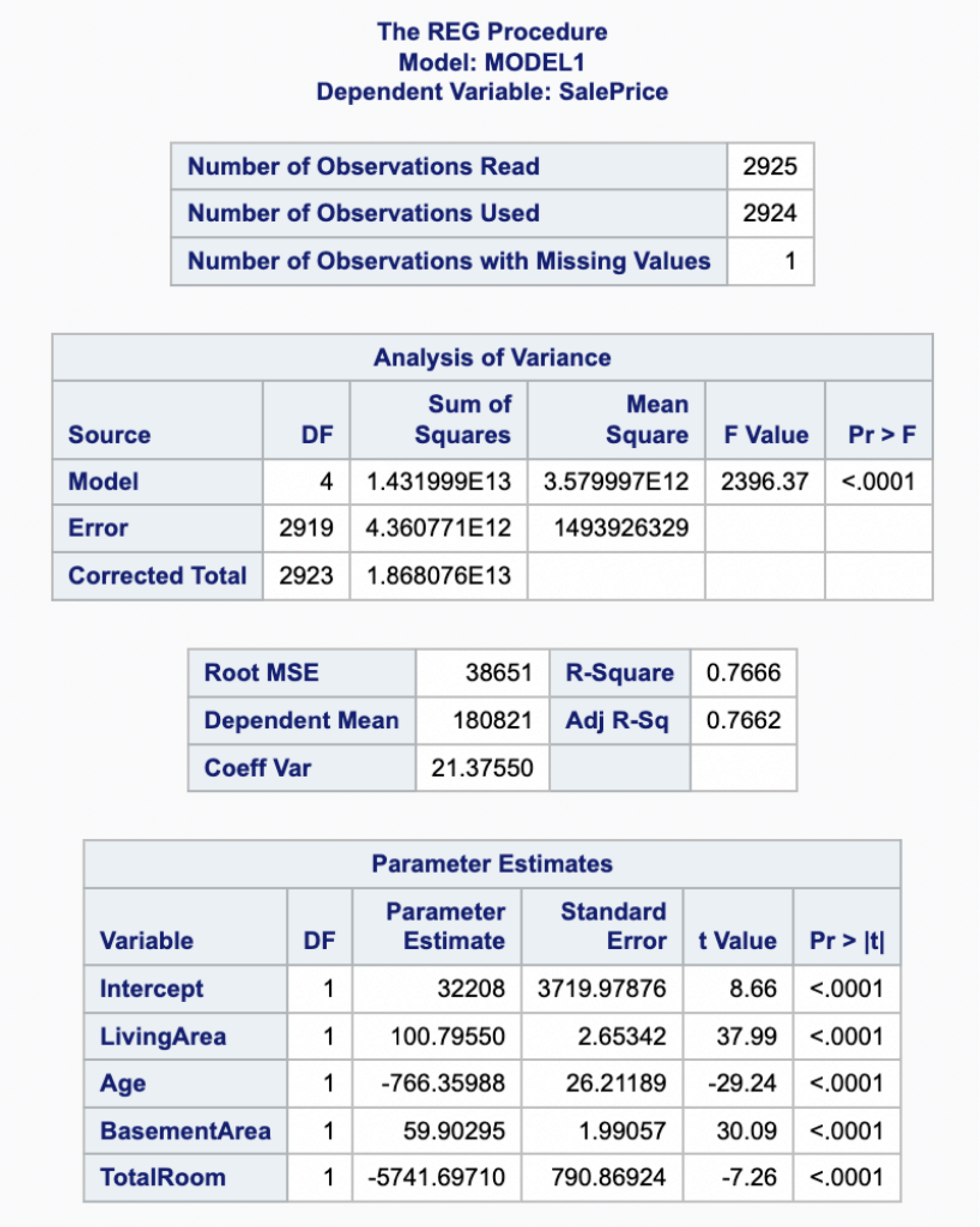
Living Area

ii.

Age

(e)

In addition to Living Area and Age, add two additional explanatory variables Basement Area and Total Room into the multiple linear regression model. The SAS output is shown below.



The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2925
Number of Observations Used	2924
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1.431999E13	3.579997E12	2396.37	<.0001
Error	2919	4.360771E12	1493926329		
Corrected Total	2923	1.868076E13			

Root MSE	38651	R-Square	0.7666
Dependent Mean	180821	Adj R-Sq	0.7662
Coeff Var	21.37550		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	32208	3719.97876	8.66	<.0001
LivingArea	1	100.79550	2.65342	37.99	<.0001
Age	1	-766.35988	26.21189	-29.24	<.0001
BasementArea	1	59.90295	1.99057	30.09	<.0001
TotalRoom	1	-5741.69710	790.86924	-7.26	<.0001

Figure 2: CocoMelon

i.

How much is the reduction to the Sums of Squares for Error for adding Basement Area and Total Room to the model with Living Area and Age?

ii.

Provide the partial F-test for the significance of Basement Area and Total Room in the model with Living Area and Age. Report the null and alternative hypotheses, test statistic and p-value, and interpret the result in context.