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 year old 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?

```r
data <- read.table("BGSgirls2.txt", header = F)

colnames(data) <- c(
  "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
)

response_var <- "BMI"
explanatory_vars <- c("HT2", "HT9", "WT2", "WT9", "ST9")

correlations <- sapply(X = explanatory_vars,
                       FUN = function(var) {
                         cor(data[[var]], data[[response_var]])
                         })
correlations
```

```
##         HT2         HT9         WT2         WT9         ST9
## 0.042573733 0.236907969 0.190947873 0.545925753 0.005603061
```

```
correlation_results <- sapply(X = explanatory_vars,
                              FUN = function(var) {
                                test <- cor.test(data[[var]], data[[response_var]])
                                c("correlation" = round(test$estimate, 4), "p-value" = round(test$p.valu
                                })

correlation_results_df <- as.data.frame(t(correlation_results))
correlation_results_df <- cbind(variable = explanatory_vars, correlation_results_df)
correlation_results_df
```

```
##     variable correlation.cor p-value
## HT2      HT2          0.0426  0.7264
## HT9      HT9          0.2369  0.0483
## WT2      WT2          0.1909  0.1133
## WT9      WT9          0.5459  0.0000
## ST9      ST9          0.0056  0.9633
```

Using `cor.test` we are able to extract both the correlation coefficient as well as its associated p-value for determining significance. For something to be "significant", we evaluate it at the $\alpha = 0.05$ level, such that we are testing whether the explanatory variable in question has a correlation coefficient with the response variable (BMI) significantly different from 0 (null hypothesis, alternative being the correlation is not zero). That being said:

Significant Findings

The only significant explanatory variables are for WT9 and HT9 at the level specified above are:

WT9: Weight (kg) at 9 years - Correlation: 0.5459 - p-value: $1.018837e{-}6$ - Significant: The correlation is strong, greater in magnitude to 0.5, and the p-value is smaller than our reference threshold.

HT9: Height (cm) at 9 years - Correlation: 0.2369 - p-value: 0.0483 - Significant: The correlation is not especially strong, but the p-value is below the 0.05 threshold.

Not Significant Findings

The other explanatory variables considered had small (close to zero estimated sample) correlations, and as such are found not to meet the significance threshold.

ST9: - Correlation: 0.0056 - p-value: 0.9633 - Not Significant: The correlation is near zero and the p-value is nearly 1.

HT2: - Correlation: 0.0426 - p-value: 0.7264 - Not Significant: The correlation is near zero and the p-value is rather large.

WT2: - Correlation: 0.1909 - p-value: 0.1133 - Not Significant: While the correlation is not especially strong, and despite being a somewhat small p-value it is nonetheless larger than the comparison value of 0.05.

(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?

```
explanatory_vars <- c("HT2", "HT9", "WT2", "WT9", "ST9")
explanatory_data <- data[, explanatory_vars]
cor_matrix <- cor(explanatory_data)
n <- nrow(explanatory_data)
p_values <- matrix(NA, nrow = length(explanatory_vars), ncol = length(explanatory_vars))
rownames(p_values) <- colnames(p_values) <- explanatory_vars

# For loops are allowed
# Yipee!
for (i in 1:length(explanatory_vars)) {
  for (j in 1:length(explanatory_vars)) {
    if (i != j) {
      test <- cor.test(explanatory_data[[i]], explanatory_data[[j]])
      p_values[i, j] <- test$p.value
    }
  }
}

significant_correlations <- which(p_values < 0.05, arr.ind = TRUE)
significant_results <- data.frame(
  Var1 = rownames(p_values)[significant_correlations[, 1]],
  Var2 = colnames(p_values)[significant_correlations[, 2]],
  "correlation" = round(cor_matrix[significant_correlations], 4),
  "p-Value" = round(p_values[significant_correlations], 4)
)

significant_results <- significant_results[!duplicated(t(apply(significant_results, 1, sort))), ]
significant_results
```

```
##    Var1 Var2 correlation p.Value
## 1   HT9  HT2      0.7384  0.0000
## 2   WT2  HT2      0.6445  0.0000
## 3   WT9  HT2      0.5229  0.0000
## 4   ST9  HT2      0.3617  0.0021
## 6   WT2  HT9      0.6071  0.0000
## 7   WT9  HT9      0.7276  0.0000
## 8   ST9  HT9      0.6034  0.0000
## 11  WT9  WT2      0.6925  0.0000
## 12  ST9  WT2      0.4516  0.0001
## 16  ST9  WT9      0.4530  0.0001
```

HT9, WT9, WT2, and ST9 are strongly or moderately correlated with multiple other variables, where "strongly" or "moderately" is whether the magnitude of correlation (|Rho|) is greater than or equal to 0.35 (for moderate, 0.50 for strong). Pairwise comparisons are listed below for variables with correlations stronger in magnitude than 0.05, which is all of them. Please pardon the redundancies as Cor(HT9, HT2) = Cor(HT2, HT9):

HT2

- HT9 $r = 0.7384$, $p \approx 3.00e-13$
- WT2 $r = 0.6445$, $p \approx 1.73e-9$
- WT9 $r = 0.5229$, $p \approx 3.42e-6$
- ST9 $r = 0.3617$, $p \approx 2.09e-3$

HT9

- HT2 $r = 0.7384$, $p \approx 3.00e-13$
- WT9 $r = 0.7276$, $p \approx 9.69e-13$
- WT2 $r = 0.6071$, $p \approx 2.52e-8$
- ST9 $r = 0.6034$, $p \approx 3.23e-8$

WT2

- WT9 $r = 0.6925$, $p \approx 3.11e-11$
- HT2 $r = 0.6445$, $p \approx 1.73e-9$
- HT9 $r = 0.6071$, $p \approx 2.52e-8$
- ST9 $r = 0.4516$, $p \approx 8.71e-5$

WT9

- HT9 $r = 0.7276$, $p \approx 9.69e-13$
- WT2 $r = 0.6925$, $p \approx 3.11e-11$
- HT2 $r = 0.5229$, $p \approx 3.42e-6$
- ST9 $r = 0.4530$, $p \approx 8.22e-5$

ST9

- HT9 $r = 0.6034$, $p \approx 3.23e-8$
- WT2 $r = 0.4516$, $p \approx 8.71e-5$
- WT9 $r = 0.4530$, $p \approx 8.22e-5$
- HT2 $r = 0.3617$, $p \approx 2.09e-3$

(c)

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

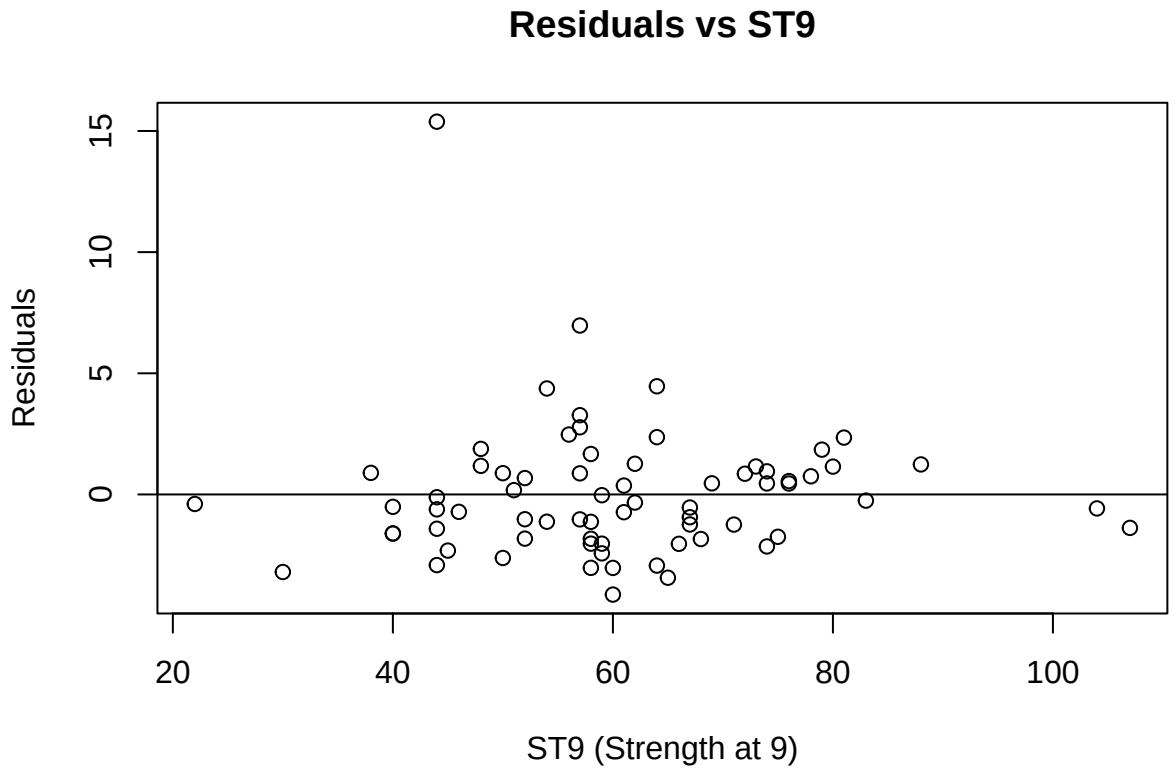
$$BMI_i = \beta_0 + \beta_1 ST9_i + \epsilon_i, \text{ for } i = 1, ..., 70$$

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

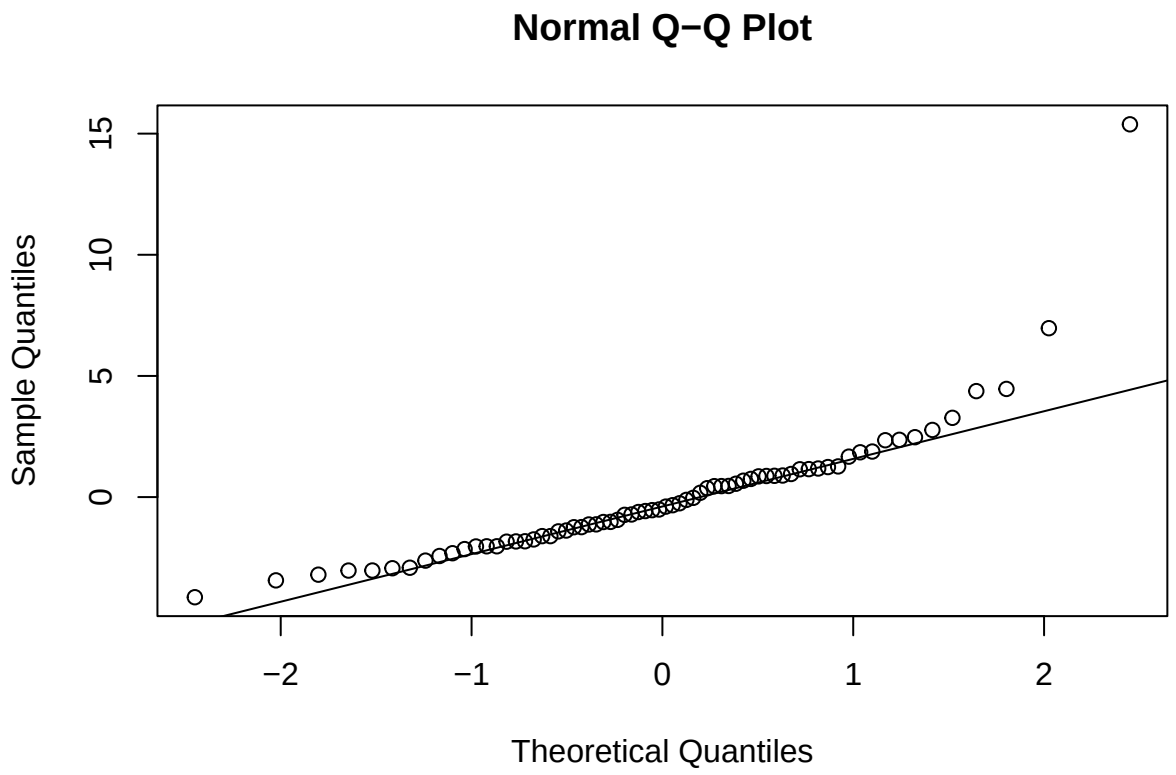
```r
model <- lm(BMI ~ ST9, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = BMI ~ ST9, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1324 -1.7138 -0.4527  0.9375 15.3840
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.471009   1.379242  15.567   <2e-16 ***
## ST9          0.001023   0.022141   0.046    0.963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.782 on 68 degrees of freedom
## Multiple R-squared:  3.139e-05,  Adjusted R-squared:  -0.01467
## F-statistic: 0.002135 on 1 and 68 DF,  p-value: 0.9633
```

```r
plot(data$ST9, resid(model), main = "Residuals vs ST9", xlab = "ST9 (Strength at 9)", ylab = "Residuals")
abline(h = 0, col = "black")
```



```r
qqnorm(resid(model))
qqline(resid(model), col = "black")
```



Least Squares Estimates:

From the regression summary: - $\hat{\beta}_0 = 21.471$: This is the estimated intercept, meaning that the predicted BMI for a girl with ST9 = 0 is 21.471. - $\hat{\beta}_1 = 0.001$: This is the estimated slope, meaning that for each one-unit increase in strength at age 9 (ST9), the predicted BMI increases by 0.001 on average.

Significance of Slope:

- Null Hypothesis (H_0): The slope $\beta_1 = 0$ (strength at age 9 has no effect on BMI at age 18).
- Alternative Hypothesis (H_a): The slope $\beta_1 \neq 0$ (strength at age 9 affects BMI at age 18).
- t-value for the slope: $t = 0.046$.
- p-value for the slope: $p = 0.963$.

Since $p = 0.963$, which is much greater than $\alpha = 0.05$, we fail to reject the null hypothesis. The slope is not significantly different from zero, meaning that strength at age 9 does not significantly predict BMI at age 18.

Residual Plots:

- Residuals vs. ST9:

  - The residuals are scattered randomly around 0, which indicates no clear pattern and suggests that the assumption of linearity is reasonable.
  - However, there is one potential outlier with a residual near 15, which could affect the results.

- Normal Q-Q Plot:
  - The residuals generally follow the red line, indicating approximate normality.
  - However, the tails deviate slightly, suggesting that the normality assumption might not hold perfectly, particularly due to the outlier.

Model Fit:

- $R^2 = 0.00003139$: This indicates that only 0.003% of the variation in BMI at age 18 is explained by strength at age 9. This is an extremely poor fit.
- Adjusted $R^2 = -0.01467$: The adjusted R^2 is negative, which further confirms that the model does not explain the data well.

(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?

```
multi_model <- lm(BMI ~ ST9 + WT9, data = data)
summary(multi_model)
```

```
##
## Call:
## lm(formula = BMI ~ ST9 + WT9, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1736 -1.2146 -0.2474  1.1231 11.2834
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.63878    1.54661   9.465 5.66e-14 ***
## ST9         -0.05552    0.01983  -2.799  0.00668 **
## WT9          0.32418    0.05151   6.293 2.72e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.222 on 67 degrees of freedom
## Multiple R-squared:  0.3715, Adjusted R-squared:  0.3528
## F-statistic:  19.8 on 2 and 67 DF,  p-value: 1.747e-07
```

Estimate of β_1 (Effect of Strength at Age 9):

- From the multiple regression model: $\hat{\beta}_1 = -0.05552$ (slope for ST9).
- From the simple regression model in part (c): $\hat{\beta}_1 = 0.001023$.

Comparison:

- The estimates for β_1 differ significantly.
- In the simple regression model, the estimate reflects the total (unadjusted) association between ST9 and BMI.
- In the multiple regression model, the estimate reflects the unique contribution of ST9 to BMI, after adjusting for WT9. The negative coefficient in the multiple regression suggests that, after accounting for weight at age 9, greater strength is associated with lower BMI.

Explanation:

- This change is expected because WT9 and ST9 are likely correlated, and including WT9 in the model adjusts for its effect. This adjustment alters the interpretation of β_1.

Is the Effect of Strength at Age 9 Significant?

- The t-value for β_1: $t = -2.799$.
- The p-value for β_1: $p = 0.00668$.

Since $p < 0.05$, the effect of strength at age 9 (ST9) is statistically significant in the multiple regression model. This indicates that strength at age 9 has a significant relationship with BMI at age 18, after adjusting for weight at age 9 (WT9).

Model Fit:

```
- $R^2 = 0.3715$: About 37.15% of the variation in BMI at age 18 is explained by the model including $\
- Adjusted $R^2 = 0.3528$: Adjusted for the number of predictors, this still shows a significant improv
```

Conclusion:

- The estimate of β_1 (effect of ST9) in the multiple regression model differs from the simple regression model due to adjustment for WT9. This is expected because WT9 and ST9 are correlated.
- In the multiple regression model, the effect of ST9 is significant ($p = 0.00668$), indicating that strength at age 9 contributes uniquely to predicting BMI at age 18, after accounting for weight at age 9.
- Adding WT9 significantly improves the model fit, as shown by the increase in R^2.

(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.

```r
full_model <- lm(BMI ~ WT2 + HT2 + WT9 + HT9 + ST9, data = data)
summary(full_model)
```

```
##
## Call:
## lm(formula = BMI ~ WT2 + HT2 + WT9 + HT9 + ST9, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0948 -1.2186 -0.2533  1.0090 10.4951
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.855335   8.781156   3.514 0.000817 ***
## WT2         -0.317779   0.278736  -1.140 0.258505
## HT2         -0.193997   0.130819  -1.483 0.142996
## WT9          0.419762   0.075211   5.581  5.2e-07 ***
## HT9          0.008057   0.096344   0.084 0.933613
## ST9         -0.044416   0.022219  -1.999 0.049853 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.14 on 64 degrees of freedom
## Multiple R-squared:  0.4431, Adjusted R-squared:  0.3996
## F-statistic: 10.19 on 5 and 64 DF,  p-value: 3.294e-07
```

$R^2 = 0.4431$:

Interpretation: 44.31% of the variability in BMI at age 18 is explained by the multiple linear regression that uses weight and height at ages 2 and 9 (WT2, HT2, WT9, HT9), and strength at age 9 (ST9) as explanatory variables. Frankly, this isn't especially great, as we can explain less than half the variability in what we are attempting to estimate.

(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.

```r
summary(full_model)
```

```
##
## Call:
## lm(formula = BMI ~ WT2 + HT2 + WT9 + HT9 + ST9, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0948 -1.2186 -0.2533  1.0090 10.4951
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.855335   8.781156   3.514 0.000817 ***
```

```
## WT2         -0.317779   0.278736  -1.140 0.258505
## HT2         -0.193997   0.130819  -1.483 0.142996
## WT9          0.419762   0.075211   5.581  5.2e-07 ***
## HT9          0.008057   0.096344   0.084 0.933613
## ST9         -0.044416   0.022219  -1.999 0.049853 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.14 on 64 degrees of freedom
## Multiple R-squared:  0.4431, Adjusted R-squared:  0.3996
## F-statistic: 10.19 on 5 and 64 DF,  p-value: 3.294e-07
```

Intercept (β_0):

- Null Hypothesis (H_0): $\beta_0 = 0$ (BMI at age 18 is 0 when all predictors are 0).
- Alternative Hypothesis (H_a): $\beta_0 \neq 0$.
- Interpretation: The intercept is significant ($p = 0.0008$), indicating that the predicted baseline BMI (when all predictors are 0) is non-zero. However, the intercept may not be meaningful in this context due to the lack of real-world relevance for predictors being zero.

Weight at Age 2 (β_1):

- Null Hypothesis (H_0): $\beta_1 = 0$ (Weight at age 2 has no effect on BMI at age 18, after accounting for other predictors).
- Alternative Hypothesis (H_a): $\beta_1 \neq 0$.
- Interpretation: The coefficient for WT2 is not significant ($p = 0.2585$), suggesting that weight at age 2 does not significantly predict BMI at age 18 after adjusting for other variables.

Height at Age 2 (β_2):

- Null Hypothesis (H_0): $\beta_2 = 0$ (Height at age 2 has no effect on BMI at age 18, after accounting for other predictors).
- Alternative Hypothesis (H_a): $\beta_2 \neq 0$.
- Interpretation: The coefficient for HT2 is not significant ($p = 0.1430$), suggesting that height at age 2 does not significantly predict BMI at age 18 after adjusting for other variables.

Weight at Age 9 (β_3):

- Null Hypothesis (H_0): $\beta_3 = 0$ (Weight at age 9 has no effect on BMI at age 18, after accounting for other predictors).
- Alternative Hypothesis (H_a): $\beta_3 \neq 0$.
- Interpretation: The coefficient for WT9 is highly significant ($p = 5.2 \times 10^{-7}$), indicating that weight at age 9 is a strong predictor of BMI at age 18, after adjusting for other predictors.

Height at Age 9 (β_4):

- Null Hypothesis (H_0): $\beta_4 = 0$ (Height at age 9 has no effect on BMI at age 18, after accounting for other predictors).
- Alternative Hypothesis (H_a): $\beta_4 \neq 0$.
- Interpretation: The coefficient for HT9 is not significant ($p = 0.9336$), suggesting that height at age 9 does not significantly predict BMI at age 18 after adjusting for other variables.

Strength at Age 9 (β_5):

- Null Hypothesis (H_0): $\beta_5 = 0$ (Strength at age 9 has no effect on BMI at age 18, after accounting for other predictors).
- Alternative Hypothesis (H_a): $\beta_5 \neq 0$.
- Interpretation: The coefficient for ST9 is marginally significant ($p = 0.0499$), indicating that strength at age 9 has a small but significant effect on BMI at age 18, after adjusting for other predictors. The negative sign suggests that greater strength at age 9 is associated with a slightly lower BMI at age 18.

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.

Intercept (b_0): The intercept represents the expected Sale Price of a house when both Living Area (X_1) and Age (X_2) are 0. From the output, we estimate $b_0 = 66,464$. However, I cannot reasonably say that having Living Area and Age both set to input 0 makes sense in interpretation.

Living Area (b_1): The parameter $b_1 = 102.62349$ represents the expected change in Sale Price for each additional square foot of Living Area, holding Age constant. This suggests that, on average, a one-square-foot increase in Living Area is associated with an increase of approximately \$102.62 in Sale Price when holding Age constant.

Age (b_2): The parameter $b_2 = -1072.88363$ represents the expected change in Sale Price for each additional year of Age, holding Living Area constant. This indicates that, on average, each additional year in the house's age is associated with a decrease of approximately \$1,072.88 in Sale Price when holding Living Area constant.

(b)

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

An R^2 of 0.6865 means that approximately 68.65% of the variability in Sale Price is explained by the linear regression model that includes Living Area and Age as explanatory variables (while also including an intercept term).

(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.

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

$$H_a : \text{At least one } \beta_i \neq 0 \text{ (for } i = 1, 2).$$

From the ANOVA table: - Model Sum of Squares (SSM): $1.283227e13$ - Error Sum of Squares (SSE): $5.85885e12$ - Total Sum of Squares (SST): $1.869112e13$ - Degrees of Freedom for the Model: 2 - Degrees of Freedom for Error: 2922

The F-test statistic is given by:

$$F = \frac{\text{MSM}}{\text{MSE}}$$

$$\text{MSM} = \frac{\text{SSM}}{\text{DF}_{\text{Model}}}, \quad \text{MSE} = \frac{\text{SSE}}{\text{DF}_{\text{Error}}}$$

$\text{MSM} = \frac{1.283227 \times 10^{13}}{2} = 6.416137 \times 10^{12}$ $\text{MSE} = \frac{5.85885 \times 10^{12}}{2922} = 2005082271$

$$F = \frac{6.416137 \times 10^{12}}{2005082271} \approx 3199.94$$

From the output, the p-value is reported as < 0.0001. This indicates overwhelming evidence in favor of rejecting the null hypothesis, which is evidence in favor of the alternative hypothesis (as stated above). Interpretation: The F-test interpretation stated previously is evidence that at least one of the predictors

(Living Area and Age) are significantly different from 0 in their estimated slopes, which is to say that at least one of the explanatory variables is able to improve the amount of variability we are able to explain in Sales Price by using a multiple linear regression.

(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

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

Parameter Estimate ($\hat{\beta}_1$): 102.62349 Standard Error (SE): 1.74084

$$\text{Test Statistic} = t = \frac{\hat{\beta}_1}{\text{SE}} = \frac{102.62349}{1.74084} \approx 58.95$$

From the output, the p-value < 0.0001 for Living Area, indicating overwhelming evidence in favor of rejecting the null hypothesis. Interpretation: The t-test for Living Area indicates that it is a statistically significant predictor of Sale Price (we estimate the slope of the Living Area parameter to be not equal to zero). This is to say that, after controlling for Age, there is strong evidence that Living Area is associated with changes in Sale Price, specifically that for each additional square foot of Living Area, the Sale Price is expected to increase by approximately \$102.62.

ii.

Age

$$H_0 : \beta_2 = 0$$

$$H_a : \beta_2 \neq 0$$

Parameter Estimate ($\hat{\beta}_2$): -1072.88363 Standard Error (SE): 28.21248

$$\text{Test Statistic} = t = \frac{\hat{\beta}_2}{\text{SE}} = \frac{-1072.88363}{28.21248} \approx -38.03$$

From the output, the p-value < 0.0001 for Age, indicating overwhelming evidence in favor of rejecting the null hypothesis. Interpretation: The t-test for Age indicates that it is a statistically significant predictor of Sale Price (we estimate the slope of the Age parameter to be not equal to zero). This is to say that, after controlling for Living Area, there is strong evidence that Age is associated with changes in Sale Price, specifically that for each additional year of Age, the Sale Price is expected to decrease by approximately \$1,072.88.

(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?

SSE for the initial model (Living Area and Age): $5.85885e12$ SSE for the updated model (Living Area, Age, Basement Area, Total Room): $4.36077e12$

Reduction in SSE = SSE (initial model)$-$SSE (updated model) = $5.85885\times10^{12}-4.36077\times10^{12} = 1.49808\times10^{12}$

By adding Basement Area and Total Room to the model, the Sum of Squares for Error was reduced by $1.49808e12$.

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.

$$H_0 : \beta_{\text{BasementArea}} = \beta_{\text{TotalRoom}} = 0$$

$$H_a : \text{At least one of } \beta_{\text{BasementArea}} \text{ or } \beta_{\text{TotalRoom}} \neq 0$$

Our "m" is 2 as we add 2 additional explanatory variables. Reduction in SSE: $1.49808e12$ (calculated in part i) Mean Squared Error (MSE) of the "full" model: 1493926329

$$F = \frac{\frac{SSE_{reduction}}{m}}{MSE_{full}} = \frac{\left(\frac{1.49808\times10^{12}}{2}\right)}{1493926329} \approx 5014.58$$

Using the output and the extremely large F-value of 5014.58, the p-value is reported as < 0.0001. This indicates overwhelming evidence in favor of rejecting the the null hypothesis. Interpretation: The partial F-test shows that adding Basement Area and Total Room significantly improves the model for predicting Sale Price, beyond what is explained by Living Area and Age alone.