STAT 500 Homework 10

Vu Thi-Hong-Ha; NetID: 851924086

November 6, 2020

1 Question 1

(a) The two equations are:

$\hat{Y}_i = 671.82694 + 0.71675 * engine + 26.96661 * cylinder + (-15.36376) * cityMPG + (-5.43717) * gears + (11.59035) * \mathbb{1}(intake = 1) + (-4.38453) * \mathbb{1}(intake = 1) * cylinder$

If $intake = 1$: $\hat{Y}_i = 683.4173 + 0.71675 * engine + 22.58208 * cylinder + (-15.36376) * cityMPG + (-5.43717) * gears$

If $intake = 0$: $\hat{Y}_i = 671.82694 + 0.71675 * engine + 26.96661 * cylinder + (-15.36376) * cityMPG + (-5.43717) * gears$

The differences are in the intercept and the coefficient for cylinder. The intercepts are different by the amount of the coefficient for intake, and the coefficients for cylinder are different by the amount of coefficient for the interaction between cylinder and intake.

(b) The null and alternative hypothesis for the t-test for interaction between cylinder and intake is:

$H_0 : \beta_6 = 0;$

$H_a : \beta_6 \neq 0$

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	671.82694	86.48882	7.77	<.0001	501.24229	842.41159
engine	1	0.71675	5.53173	0.13	0.8970	-10.19367	11.62717
cylinder	1	26.96661	11.99347	2.25	0.0257	3.31152	50.62171
cityMPG	1	-15.36376	0.68368	-22.47	<.0001	-16.71220	-14.01533
gears	1	-5.43717	1.63596	-3.32	0.0011	-8.66383	-2.21050
intake	1	11.59035	84.99299	0.14	0.8917	-156.04403	179.22472
cylinderxintake	1	-4.38453	10.81762	-0.41	0.6857	-25.72046	16.95140

The t-test statistic is -0.41 with a p-value of 0.6857. At a significance level of 0.05, we fail to reject the null that the coefficient for the interaction between cylinder and intake is 0. We conclude that the interaction between cylinder and intake is not statistically significant in a model that includes Engine, Cylinder, City MPG and Gears and Intake.

(c) In our model, it includes Engine, Cylinder, City MPG, Gears, Intake and the interaction between Intake and Cylinder. Therefore, we should not test for significance of Intake and Cylinder because the component variable is included by the interaction term, and we Cannot separate significance of component variable from its interaction term.

2 Question 2

(a)

Pearson Correlation Coefficients, N = 70
Prob > |r| under H0: Rho=0

	bmi	ht2	ht9	wt2	wt9	st9
bmi	1.00000	0.04257 0.7264	0.23691 0.0483	0.19095 0.1133	0.54593 <.0001	0.00560 0.9633
ht2	0.04257 0.7264	1.00000	0.73836 <.0001	0.64455 <.0001	0.52293 <.0001	0.36172 0.0021
ht9	0.23691 0.0483	0.73836 <.0001	1.00000	0.60712 <.0001	0.72761 <.0001	0.60337 <.0001
wt2	0.19095 0.1133	0.64455 <.0001	0.60712 <.0001	1.00000	0.69254 <.0001	0.45158 <.0001
wt9	0.54593 <.0001	0.52293 <.0001	0.72761 <.0001	0.69254 <.0001	1.00000	0.45300 <.0001
st9	0.00560 0.9633	0.36172 0.0021	0.60337 <.0001	0.45158 <.0001	0.45300 <.0001	1.00000

The variable HT2 (pvalue = 0.0483) and WT9 ($pvalue < 0.001$) are significantly correlated with BMI at a significance level of 0.05.

(b) At a significance level of 0.05:

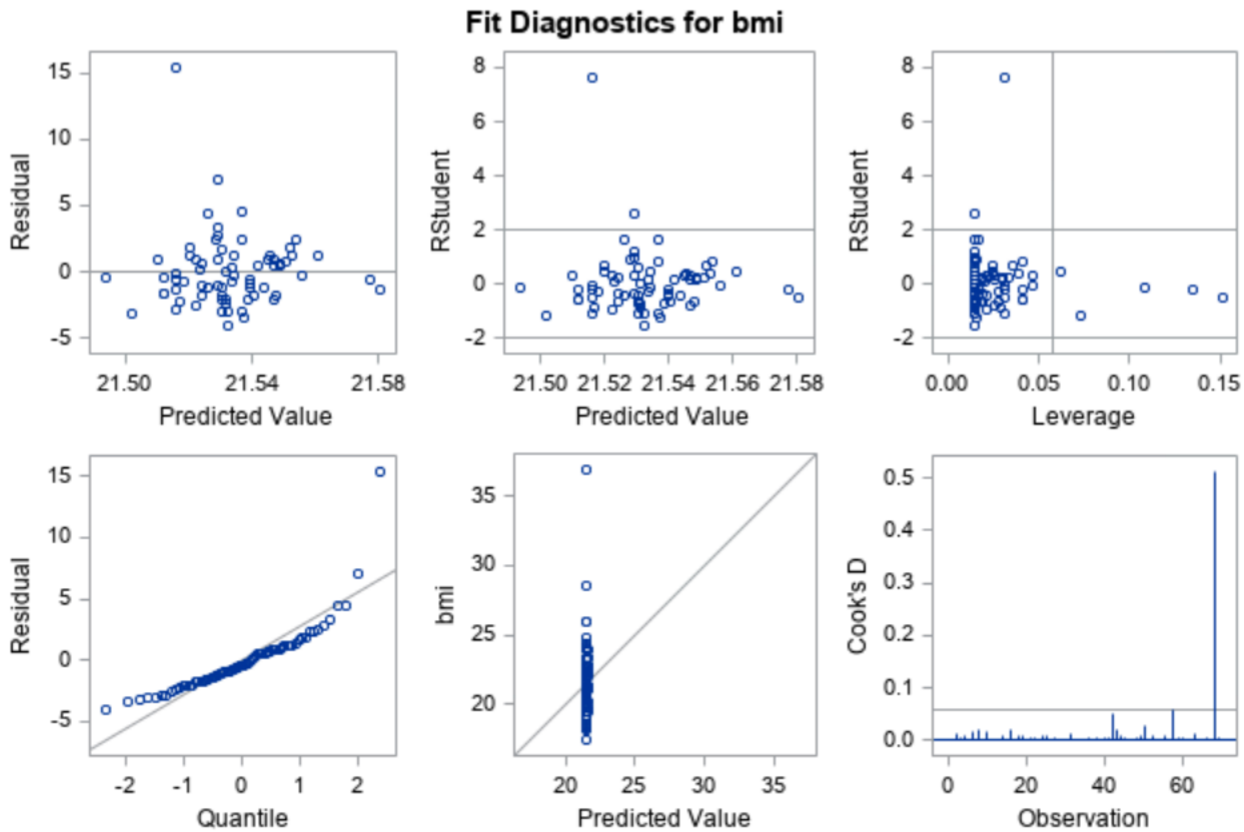
- HT2 is significantly correlated with HT9 ($pvalue < 0.001$), WT2 ($pvalue < 0.001$), WT9 ($pvalue < 0.001$), and ST9 ($pvalue = 0.0021$);
- HT9 is significantly correlated with HT2 ($pvalue < 0.001$), WT2 ($pvalue < 0.001$), WT9 ($pvalue < 0.001$), and ST9 ($pvalue < 0.001$);
- WT2 is significantly correlated with HT2 ($pvalue < 0.001$), HT9 ($pvalue < 0.001$), WT9 ($pvalue < 0.001$), and ST9 ($pvalue < 0.001$);
- WT9 is significantly correlated with HT2 ($pvalue < 0.001$), HT9 ($pvalue < 0.001$), WT2 ($pvalue < 0.001$), and

ST9 ($pvalue < 0.001$);

- ST9 is significantly correlated with HT2 ($pvalue = 0.0021$), HT9 ($pvalue < 0.001$), and WT2 ($pvalue < 0.001$).

(c)

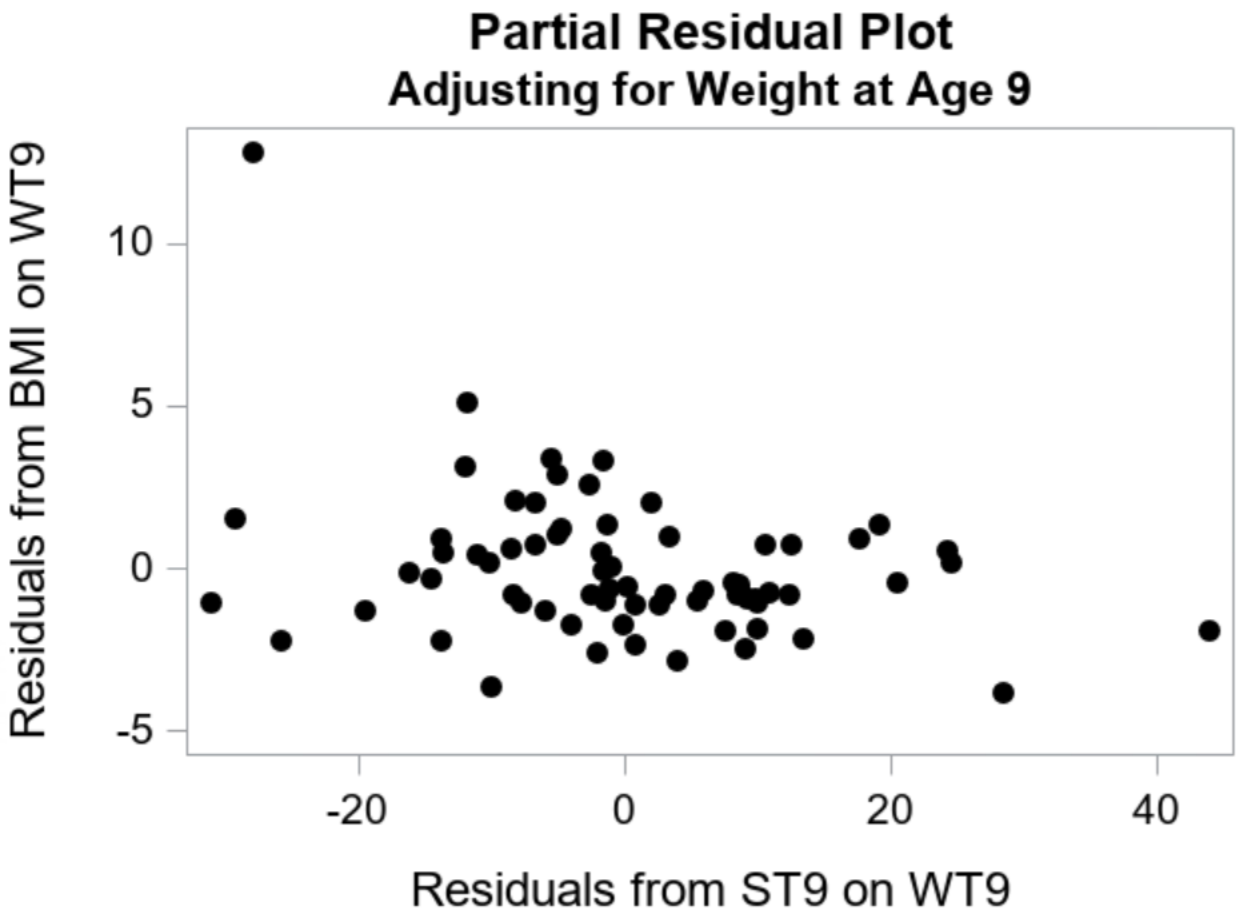
Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	21.47101	1.37924	15.57	<.0001	18.71877	24.22324
st9	1	0.00102	0.02214	0.05	0.9633	-0.04316	0.04520



The estimated intercept is 21.47101 and the estimated slope is 0.00102. The t-test statistic for the slope is 0.05 with a p-value of 0.9633. At a significance level of 0.05, we fail to reject the null that the slope is zero.

From the Predicted Value vs. Residual plot, it does not seem there is a linear relationship. There is no clear pattern of data points, indicating homogeneous variance. From the Predicted Value vs. RStudent plot and the Leverage vs. RStudent plot, it looks like there are a few outliers. The q-q plot implies that the data is right skewed.

(d)



The data points seem to follow a curve, indicating that after adjusting out the effects of WT9, BMI and ST9 may have some nonlinear relationship, and we may have to transform the data. There seem to have some outliers (point with residual from BMI on WT9 ¿ 10 and that with residual from ST9 on WT9 ¿ 40).

(e)

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2.5042E-15	0.26358	-0.00	1.0000
rst9wt9	Residual	1	-0.05552	0.01969	-2.82	0.0063

The estimated slope is -0.05552 with standard error 0.01969.

(f)

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	14.63878	1.54661	9.47	<.0001
wt9	1	0.32418	0.05151	6.29	<.0001
st9	1	-0.05552	0.01983	-2.80	0.0067

The estimated coefficient for ST9 is -0.05552 with standard error 0.01969, which is not the same as that in part (c). I expected these estimates to be different because in the current model (part f), we have more explanatory variable.

(g) The estimated slopes in part (e) and (f) are the same, and both of the standard errors are also the same.

(h) $R^2 = 0.4431$. 44.31% of the variation in BMI values at age 18 can be attributed changes in the conditional means for BMI as the five explanatory variables vary across girls.

(i)

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	30.85533	8.78116	3.51	0.0008
ht2	1	-0.19400	0.13082	-1.48	0.1430
wt2	1	-0.31778	0.27874	-1.14	0.2585
ht9	1	0.00806	0.09634	0.08	0.9336
wt9	1	0.41976	0.07521	5.58	<.0001
st9	1	-0.04442	0.02222	-2.00	0.0499

$\hat{\beta}_0 = 30.85533$ with standard error 8.78116. T test for null hypothesis $H_0 : \beta_0 = 0$ against alternative hypothesis $H_a : \beta_0 \neq 0$. T test statistic is 3.51 with p-value 0.0008. At a significance level of 0.05, we reject the null that the intercept is 0.

$\hat{\beta}_1 = -0.194$ with standard error 0.13082. T test for null hypothesis $H_0 : \beta_1 = 0$ against alternative hypothesis $H_a : \beta_1 \neq 0$. T test statistic is -1.48 with p-value 0.143. At a significance level of 0.05, we fail to reject the null that β_1 is 0.

$\hat{\beta}_2 = -0.31778$ with standard error 0.27874. T test for null hypothesis $H_0 : \beta_2 = 0$ against alternative hypothesis $H_a : \beta_2 \neq 0$. T test statistic is -1.14 with p-value 0.2585. At a significance level of 0.05, we fail to reject the null that β_2 is 0.

$\hat{\beta}_3 = 0.00806$ with standard error 0.09634. T test for null hypothesis $H_0 : \beta_3 = 0$ against alternative hypothesis $H_a : \beta_3 \neq 0$. T test statistic is 0.08 with p-value 0.9336. At a significance level of 0.05, we fail to reject the null that β_3 is 0.

$\hat{\beta_4} = 0.41976$ with standard error 0.07521. T test for null hypothesis $H_0 : \beta_4 = 0$ against alternative hypothesis $H_a : \beta_4 \neq 0$. T test statistic is 5.58 with p-value $< .0001$. At a significance level of 0.05, we reject the null that β_4 is 0.

$\hat{\beta_5} = -0.04442$ with standard error 0.02222. T test for null hypothesis $H_0 : \beta_5 = 0$ against alternative hypothesis $H_a : \beta_5 \neq 0$. T test statistic is -2.00 with p-value 0.0499. At a significance level of 0.05, we reject the null that β_5 is 0.

3 Question 3

(a) We are considering the model $\hat{Y_i} = \beta_0 + \beta_1 * LivingArea + \beta_2 * Age$

The intercept β_0 is the sale price when the living area is 0 and the age of the place is 0. The coefficient β_1 represents the change in sale price when the living area increases by 1 sqrt foot. β_2 represents the change in sale price when the age of the place increases by 1 year.

(b) $R^2 = 0.6865$, indicating that 68.65% of the variation in price can be explained by the multiple linear regression model with both living area and age.

(c)

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			

F test tests for the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ against the alternative hypothesis H_a : At least one $\beta_i \neq 0$ where i = 1 or 2. The F value is 3199.94 with a pvalue $< .0001$. Therefore, at a significance level of 0.05, we reject the null that both of the coefficients are zero. We conclude that at least one of the coefficients is nonzero.

(d)

For living area variable: T test tests for the null hypothesis $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_a : \beta_1 \neq 0$. The t value is 58.95 with a pvalue $< .0001$. Therefore, at a significance level of 0.05, we reject the null that the coefficient of the living area variable is zero.

For age variable: T test tests for the null hypothesis $H_0 : \beta_2 = 0$ against the alternative hypothesis $H_a : \beta_2 \neq 0$. The t value is -38.03 with a pvalue $< .0001$. Therefore, at a significance level of 0.05, we reject the null that the coefficient of the age variable is zero.

(e)

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			

The sum of square for error when adding basement area and total room to the model decreased by 1.498079E12.

Partial F test:

$H_0 : \beta_j = 0$ for the variables of basement area and total room.

$H_a :$ At least one $\beta_j \neq 0$ for the variables of basement area and total room.

Test statistic:

$$F = \frac{(SSE_{r.model} - SSE_{f.model})/m}{MSE_{f.model}} = \frac{(5.85885E12 - 4.360771E12)/2}{1493926329} = 501.3899$$

$F_{m,n-(k+1),1-\alpha} = F_{2,2919,0.95}$

$pvalue < 0.0001$. At a significance level of 0.05, we reject the null that the coefficients for the variables of basement area and total room