Elliot Weiss
BSTT 401
05/06/2020

Mini Project – Final Report

Introduction:

Thyroid stimulating hormone, or TSH, is a hormone secreted by the pituitary gland. TSH regulates the function of the thyroid gland, which produces hormones T_3 and T_4 in order to control metabolism of tissues in the body. When TSH levels are too high or too low, this can affect the secretion of T_3 and T_4 by the thyroid gland, resulting in hypothyroidism or hyperthyroidism, and thus affecting the regulation of metabolism in the body. Both disorders of the thyroid gland are found to be common as age increases (Hershman, 2019). In some studies, high TSH levels have been associated with an increased risk of thyroid cancer (Huang, et. al, 2017). In this report, TSH level was selected as the outcome variable, and analyzed with a multitude of predictors, including: sex, age, BMI, poverty income ratio (PIR), ethnicity, education level, smoking status, and metals measured from blood and urine samples. All predictors in SAS are numerical variables, except for ethnicity, which is a character variable. For the purpose of this analysis, ethnicity was converted from a character variable to a numerical variable.

Univariate Analysis:

Table 1. Descriptive Statistics for Outcome Variable: TSH

Outcome Variable: TSH					
Sample Size	**Mean**	**Standard Deviation**	**Variance**	**Minimum**	**Maximum**
3747	1.9301	1.9513	3.8075	0.5050	69.8360

With a sample size of 3747 individuals, TSH levels were measured. The average TSH level measured in the blood in this sample was 1.93 uIU/mL, with a standard deviation of 1.95 uIU/mL. Variance was reported to be 3.80 uIU/mL. Both standard deviation and variance appear to be small and close to the mean. The lowest recorded TSH level was 0.5 uIU/mL, whereas the highest recorded TSH level was 69.84 uIU/mL.

Bivariate Analysis:

Table 2. Bivariate Correlation Analysis

Pearson Correlation Coefficients, N = 3747 Prob > |r| under H0: Rho=0							
	Sex	**Age**	**Ethnicity**	**BMI**	**EDUC**	**PIR**	**SMOKER**
TSH	**0.02488** **0.1279**	**0.10150** **<.0001**	**0.08711** **<.0001**	**0.04314** **0.0083**	**-0.01157** **0.4788**	**0.02426** **0.1550**	**-0.03983** **0.0147**
	Cadmium	**Lead**	**Barium**	**Arsenic**	**Tungsten**	**Thallium**	**Uranium**
TSH	**-0.03322** **0.0420**	**0.01984** **0.2247**	**0.00743** **0.6493**	**-0.00124** **0.9395**	**0.00837** **0.6087**	**-0.00119** **0.9422**	**0.00919** **0.5739**

Upon bivariate analysis, correlation between the main outcome variable and predictors was calculated using SAS. TSH was highly correlated (alpha level = 0.05) with age (p-value <0.0001), ethnicity (p-value <0.0001), BMI (p-value = 0.0083), smoking status (p-value = 0.0147), and the metal Cadmium (p-value = 0.0420). TSH was weakly correlated with sex (p-value = 0.1279) and PIR (p-value = 0.1550). TSH was not correlated with education level, Lead, Barium, Arsenic, Tungsten, Thallium, or Uranium. For all predictors and the outcome variable, Pearson Correlation coefficients were very small and close to zero, indicating a non-linear relationship between TSH and these predictors.

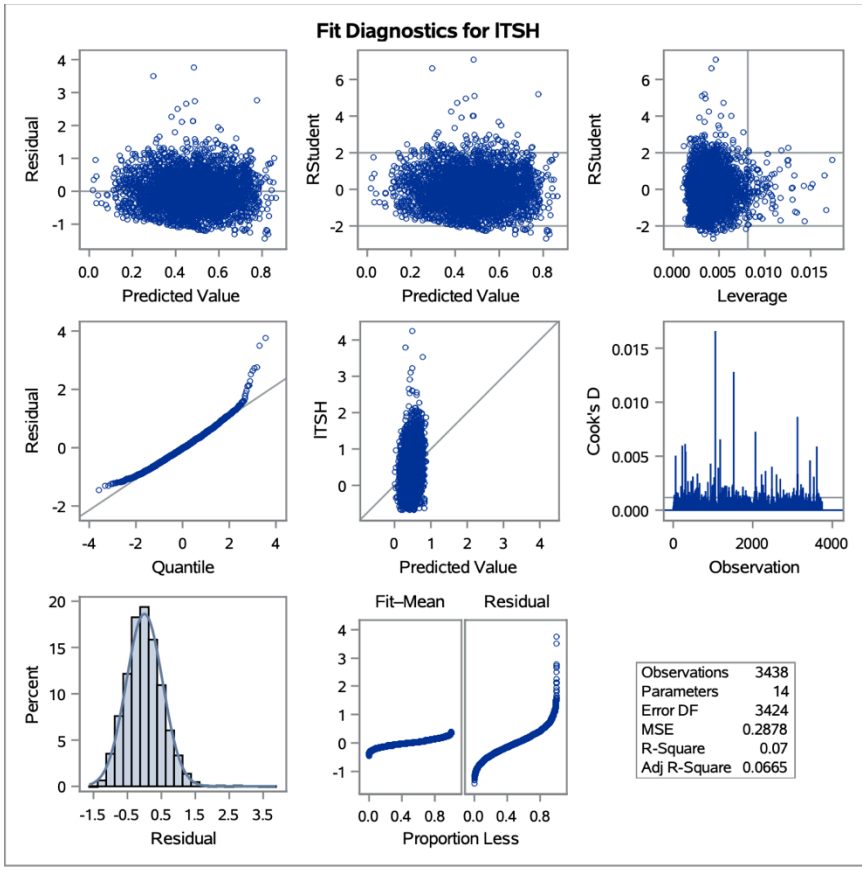
Multivariate Analysis:

For multivariate analysis, log transformations were recommended for the outcome variable, TSH, and the metals.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	74.15447	5.70419	19.82	<.0001
Error	3424	985.41556	0.28780		
Corrected Total	3437	1059.57003			

Root MSE	0.53647	R-Square	0.0700
Dependent Mean	0.47609	Adj R-Sq	0.0665
Coeff Var	112.68251		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type II SS
Intercept	1	-0.23800	0.22823	-1.04	0.2971	0.31294
FEMALE	1	0.02437	0.02085	1.17	0.2424	0.39343
AGE	1	0.00507	0.00063008	8.05	<.0001	18.63690
ETH	1	0.07920	0.00798	9.93	<.0001	28.35833
BMI	1	0.00684	0.00143	4.79	<.0001	6.61412
PIR	1	-0.00239	0.00601	-0.40	0.6911	0.04546
SMOKER	1	-0.02339	0.03153	-0.74	0.4582	0.15838
lCAD	1	-0.05182	0.01687	-3.07	0.0021	2.71444
lLead	1	0.01735	0.01734	1.00	0.3170	0.28826
lBAR	1	-0.01834	0.01072	-1.71	0.0871	0.84286
lARS	1	0.00642	0.01001	0.64	0.5214	0.11835
lTUNG	1	0.01790	0.01160	1.54	0.1227	0.68604
lTHAL	1	-0.02383	0.01897	-1.26	0.2091	0.45431
lURAN	1	0.01271	0.01142	1.11	0.2659	0.35632



Upon multivariate analysis with recommended log transformations on TSH and the metal predictors, the full model was statistically significant with a p-value < 0.0001 according to the ANOVA table. However, according to parameter estimates, the only significant predictors in the model were age, ethnicity, BMI, and logCadmium. Other weakly significant predictors in the model include logBarium and logTungsten. For the full model, goodness of fit statistics are very weak, with $R^2 = 0.07$ and RMSE being greater than the dependent mean.

According to residual diagnostics on the right, in the (1,2) plot, residuals appear to be well-scattered and homoscedastic. Normality can be assumed according to the (2,1) plot, as most of the residuals follow the 45 degree line. We can assume normal distribution according to the (3,1) plot, as the histogram appears to be unimodal and symmetric. Assuming that subjects in this analysis are independent of each other, we can conclude that residuals are iid $N(0, \sigma)$.

In order to select which variables best fit the model, backward, stepwise and C(p) selection methods were implemented using SAS. For backward selection, significance level for staying was set to alpha = 0.15. Similarly, for stepwise selection, significance level for entry and staying were both set to alpha = 0.15, respectively. For C(p) method, the best three models were displayed.

Results:

The best predictive model with consistency across all selection methods was:

$\hat{LogTSH}$ = Age + Ethnicity + BMI + logCadmium + logBarium + logTungsten

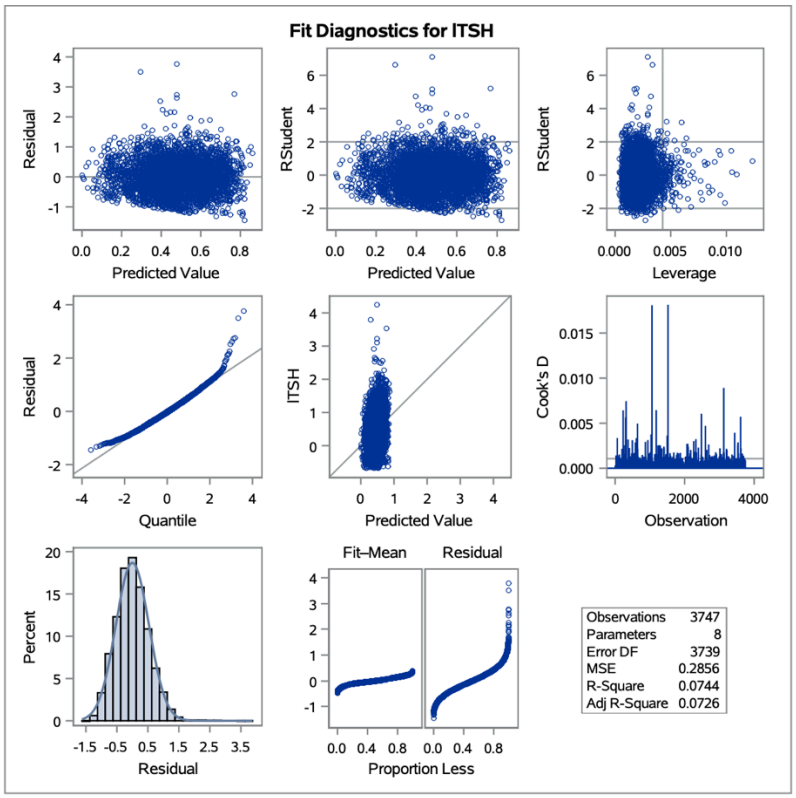
Because ethnicity was included in the best predictive model, dummy variables for each race category were inserted into the model to further understand which ethnicity best predicts the variability in TSH levels. The dummy variables were: BLAM (African American), MXAM (Mexican American), OHAM (other), and WHAM (White American). Using the C(p) selection method, the best three predictive models were presented. The best predictive model with the highest R^2 value and lowest C(p) – p value was:

$\hat{LogTSH}$ = Age + Ethnicity + BMI + logCadmium + logBarium + logTungsten + MXAM.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	85.81643	12.25949	42.92	<.0001
Error	3739	1068.00532	0.28564		
Corrected Total	3746	1153.82176			

Root MSE	0.53445	R-Square	0.0744
Dependent Mean	0.47819	Adj R-Sq	0.0726
Coeff Var	111.76502		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type II SS
Intercept	1	-0.22700	0.12793	-1.77	0.0761	0.89937
AGE	1	0.00577	0.00050711	11.38	<.0001	36.96608
ETH	1	0.08616	0.00785	10.98	<.0001	34.43377
BMI	1	0.00707	0.00136	5.21	<.0001	7.76798
ICAD	1	-0.04688	0.01138	-4.12	<.0001	4.84476
IBAR	1	-0.01627	0.00977	-1.67	0.0959	0.79227
ITUNG	1	0.02093	0.01053	1.99	0.0470	1.12802
MXAM	1	0.07181	0.02475	2.90	0.0037	2.40451



The final predictive model equation is:

$\hat{LogTSH}$ = -0.227 + 0.0058(Age) + 0.0862(Ethnicity) + 0.0071(BMI) – 0.0469(logCadmium) – 0.0163(logBarium) + 0.0209(logTungsten) + 0.0718(MXAM)

According to residual diagnostics on the right, in the (1,2) plot, residuals appear to be well-scattered and homoscedastic. Normality can be assumed according to the (2,1) plot, as most of the residuals follow the 45 degree line. We can assume normal distribution according to the (3,1) plot, as the histogram appears to be unimodal and symmetric. Assuming that subjects in this analysis are independent of each other, we can conclude that residuals are iid N(0, σ).

In the final predictive model, all predictors but one are significant in the model. These significant predictors are: age (p-value <0.0001), ethnicity (p-value <0.0001), BMI (p-value <0.0001), logCadmium (p-value <0.0001), logTungsten (p-value = 0.0470), and the dummy variable MXAM (p-value = 0.0037). LogBarium was the only predictor that was not significant in the model, but was still kept because it was weakly significant with a p-value of 0.0959.

Discussion:

In order to fully understand the final predictive model, we must explain the effect of significant predictors on the outcome variable. However, these interpretations become more complicated due to the implementation of log transformations on the outcome variable and some of the predictors. Therefore, simplified interpretations of these effects are as follows:

$\hat{LogTSH}$ = -0.227 + 0.0058(Age) + 0.0862(Ethnicity) + 0.0071(BMI) – 0.0469(logCadmium) – 0.0163(logBarium) + 0.0209(logTungsten) + 0.0718(MXAM)

As **age** increases by 1 unit, the average change in logTSH is an increase of 0.0058 uIU/mL, holding all other predictors constant. As **ethnicity**, increases by 1 unit, the average change in logTSh is an increase of 0.0862 uIU/mL, holding all other predictors fixed. As **BMI** increases by 1 unit, average logTSH increases by 0.0071

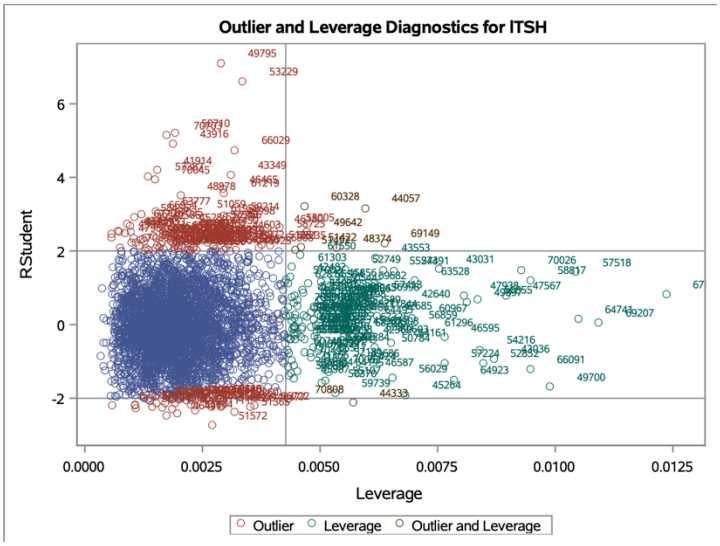
uIU/mL, holding all other predictors fixed. As **logCadmium** increases by 1 unit, the average change in logTSH is a decrease of 0.0469 uIU/mL, holding all other predictors fixed. As **logTungsten** increase by 1 unit, the average change in logTSH is an increase of 0.0209 uIU/mL, holding all other predictors fixed. Finally, dummy variable **MXAM** shows that Mexican Americans contribute an increase of 0.0718 uIU/mL to the average change of logTSH, holding all other predictors fixed.

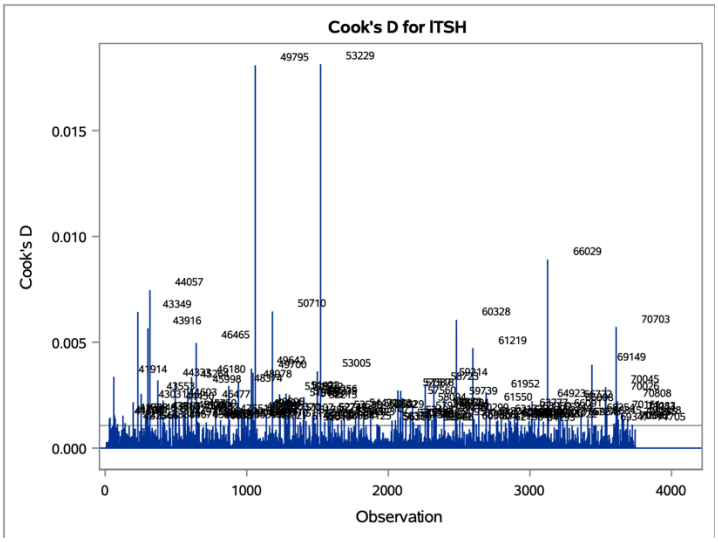
According to Type II sum of squares (SS) from the parameter estimates table in the results section, age, ethnicity, BMI, and logCadmium are the top four predictors that contribute the most in predicting the outcome variable. Age, which has the greatest Type II SS of 36.97, is consistent with the literature in predicting levels of TSH in the blood. As mentioned in the introduction, disorders of the thyroid gland are common as age increases (Hershman, 2019). Next, ethnicity has the second highest Type II SS of 34.43. According to one study which assessed demographic characteristics in newborns in order to predict TSH levels and hypothyroidism, ethnicity was determined to have a statistically significant impact on these outcomes (Heather, et. al, 2019). Next, BMI had a Type II SS of 7.77. Although it does not have as great of an impact on predicting TSH compared to age and ethnicity, BMI can be an important indicator of a thyroid disorder. TSH regulates the functionality in the thyroid gland, which controls metabolic processes in the tissues of the body. Someone with a high BMI may have issues with metabolism and therefore, could have hypothyroidism, which produces less T_3 and T_4, thus signaling the body to produce more TSH. Similarly, someone with a low BMI may have increased metabolism and therefore, could have hyperthyroidism, producing more T_3 and T_4, thus signaling the body to produce less TSH. Finally, logCadmium had a Type II SS of 4.84. Cadmium has been known to greatly affect the levels of hormones secreted by the pituitary gland. In a study which assessed the impact of cadmium levels in drinking water on TSH levels in rats, Cadmium displayed a dose-dependent response, where increasing levels of Cadmium increased TSH levels in the plasma (Lafuente, et. al, 2003).

Table 3.

Collinearity Estimates		
Variable	**Tolerance (Tol)**	**Variance Inflation Factor (VIF)**
Age	0.96418	1.03715
Ethnicity	0.80012	1.24981
BMI	0.96348	1.03790
logCadmium	0.94975	1.05291
logBarium	0.88536	1.12949
logTungsten	0.97062	1.03027
MXAM	0.88294	1.13258

According to table 3 above, none of the predictors have a tolerance less than the recommended strict cutoff value of 0.2, or a VIF greater than the recommended strict cutoff value of 5. Therefore, no multicollinearity is present in this regression and there is no inflation of standard errors or regression coefficients.





According to outlier and leverage diagnostics on the left, there are numerous candidates for outliers in this dataset. However, according to the Cook's Distance plot on the right, there are two greatly influential observations, indicated by respondent sequence numbers 49795 and 53229. These extreme observations may simply be attributed to typos or improper data collection procedures, but could be actual extreme observations. In this case, deletion of these observations may not have a significant effect on the regression, due to the large sample size.

Limitations:

With every analysis, there are limitations. There could be information bias from data collection, thus leading to misclassification bias. Additionally, sex was not determined to be a significant predictor in this analysis. In general, it is known that there are hormonal differences between males and females, so it was surprising to see that sex was not a significant predictor in this analysis. Lastly, measurements of metals from the blood may not be entirely reflective of the actual levels in the tissue.

There are also issues with the actual results of the regression. In terms of goodness of fit statistics, R^2 was extremely small, which is common with public health data. Our R^2 indicates that our predictors can only explain about 7.5% of the variability in our outcome variable. Additionally, RMSE was greater than the dependent mean. Even though our model is significant and most of the predictors are significant at alpha = 0.05, our goodness of fit statistics do not meet the criteria for a good regression, even when residuals are iid N(0, σ).

Conclusion:

In order to efficiently predict TSH levels in the body, one should consider age, ethnicity, BMI, and measure for levels of Cadmium, Barium, and Tungsten in the blood. Together, these factors are important indicators for TSH levels, which may assist in identifying disorders such as hypothyroidism or hyperthyroidism, as well as thyroid cancer. This analysis highlights the importance of ethnicity in predicting TSH levels or conditions of the thyroid. Public health disparities exist between ethnicities and should be taken into account in future analyses.

References:

Hershman, J. M. (2019, August). Overview of the Thyroid Gland - Hormonal and Metabolic Disorders. Retrieved from https://www.merckmanuals.com/home/hormonal-and-metabolic-disorders/thyroid-gland-disorders/overview-of-the-thyroid-gland?qt=thyroxine&alt=sh

Heather, N. L., Derraik, J. G., Webster, D., & Hofman, P. L. (2019). The impact of demographic factors on newborn TSH levels and congenital hypothyroidism screening. *Clinical Endocrinology, 91*(3), 456-463. doi:10.1111/cen.14044

Huang, H., Rusiecki, J., Zhao, N., Chen, Y., Ma, S., Yu, H., … Zhang, Y. (2017). Thyroid-Stimulating Hormone, Thyroid Hormones, and Risk of Papillary Thyroid Cancer: A Nested Case–Control Study. *Cancer Epidemiology Biomarkers & Prevention*, *26*(8), 1209–1218. doi: 10.1158/1055-9965.epi-16-0845

Lafuente, A., Cano, P. & Esquifino, A.I. (2003). Are cadmium effects on plasma gonadotropins, prolactin, ACTH, GH and TSH levels, dose-dependent?. *Biometals* **16,** 243–250 https://doi-org.proxy.cc.uic.edu/10.1023/A:1020658128413