1. **The following is SAS output from a simple linear regression. Fill in the XXXX’s but not the XX’s.**

Table

Description automatically generated

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SOURCE** | **DF** | **SQUARES** | **MEAN SQRS** | **F VALUE** | **Pr > F** |
| **MODEL** | - | - | - | - | 2.839\*10-7 |
| **ERROR** | - | 515.269 | - |  |  |
| **CORRECTED TOTAL** | 99 | 675.179 |  |  |  |

In R: pf(30.413,1,98,lower.tail=FALSE) = 2.839\*10-7

|  |  |  |  |
| --- | --- | --- | --- |
| R-square | Coeff var | Root MSE | Y mean |
| 0.76315 | 47.76 | 2.293 | 4.802 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | DF | Type I SS | Mean Sqr | Pr > F |
| - | - | - | - | 0.0001 |
| Source | DF | Type III SS | Mean Sqr | Pr > F |
| - | - | - | 159.91 | 0.001 |

1. **See the data in “Problem2\_HW8.txt” These data represent measurements on blood-cell counts taken from 50 individuals. We want to determine whether red-blood cell counts predict white-blood cell counts. Please analyze these data via a simple linear regression. Please articulate the assumptions you made in analyzing the data, show the appropriate diagnostic plots, and implement any remedial measures that you think are appropriate.**

For this problem, I will run a “proc reg” in SAS with the assumptions of constant variance, linear relationship between red-blood cell counts and white-blood cell counts, independent error terms, and that our model is uninfluenced by outliers.

SAS CODE:

ods graphics on;

proc reg data=Prob8;

model wbc=rbc;

run;

OUTPUT:

Table

Description automatically generated

Via simple linear regression, the p-value for our F-statistic indicates that red-blood cell counts do predict white-blood cell count. In fact, our R2 value indicated a relatively good linear fit. We will now check our assumptions:

A picture containing diagram

Description automatically generated

In this chart of diagnostic graphs, we can check some assumptions made on the data presented. First, by looking at the Residuals vs. Predicted values plot, we can say that the assumption for constant variance does hold for this data. Second, we can look at the QQ-plot to see that the does indeed show linearity since the data resembles a straight, diagonal line, well enough. We can then assume that our data has independent error terms, since by looking at the residuals vs. predicted value chart we see a cloud of points. Finally, the Cook’s D graph gives only a value of 0.25 for our largest outlier. We can safely say that our data is not influenced by outliers. Thus, our assumptions hold, and we can say assuredly that red-blood cell count predicts white-blood cell count.

1. **See the data in “Problem3\_HW8.txt” These data represent BMI and blood-cell counts taken from 100 individuals. We want to determine whether white-blood cell counts predict BMI. Please analyze these data via a simple linear regression. Please articulate the assumptions you made in analyzing the data, show the appropriate diagnostic plots, and implement any remedial measures that you think are appropriate.**

Once again, I will run a “proc reg” in SAS with the assumptions of constant variance, linear relationship between red-blood cell counts and white-blood cell counts, independent error terms, and that our model is uninfluenced by outliers.

SAS CODE:

ods graphics on;

proc reg data=Prob8b;

model bmi=wbc;

run;

OUTPUT:

Table

Description automatically generated

Graphical user interface

Description automatically generated

Via simple linear regression, the p-value for our F-statistic indicates that white-blood cell counts do predict BMI. Looking at the fit for this plot, however, our assumption of linearity does not appear to hold. We will now check our assumptions:

Diagram

Description automatically generated with medium confidence

We can check some assumptions made on the data presented using this grid of diagnostic plots. As stated above, our wbc vs bmi plot shows a lack of linearity for this data. First, by looking at the Residuals vs. Predicted values plot, we can say that the assumption for constant variance does NOT hold for this data since the spread of data points are not similar across the x-axis (y/predicted-values). Second, we can look at the QQ-plot to see that the data does NOT show linearity. This line appears to resemble a cubic polynomial; not good (although it being condensed on the line isn’t bad). We cannot assume that our data has independent error terms, since by looking at the residuals vs. predicted value chart we see a function/line instead of a cloud of points. Finally, the Cook’s D graph gives only a value of ~2.5 for our largest outlier. This point and a few others have pretty significant leverage; thus, our data is indeed influenced by outliers. As it stands, our assumptions do not hold, and we should try some transformations to reanalyze the data where they *can* fit assumptions. **Since the wbc vs bmi plot appears to be a parabola, I’ll use a root transformation on bmi to get back to linearity**:

SAS CODE:

data Prob8b;

set Prob8b;

rootbmi=sqrt(bmi);

run;

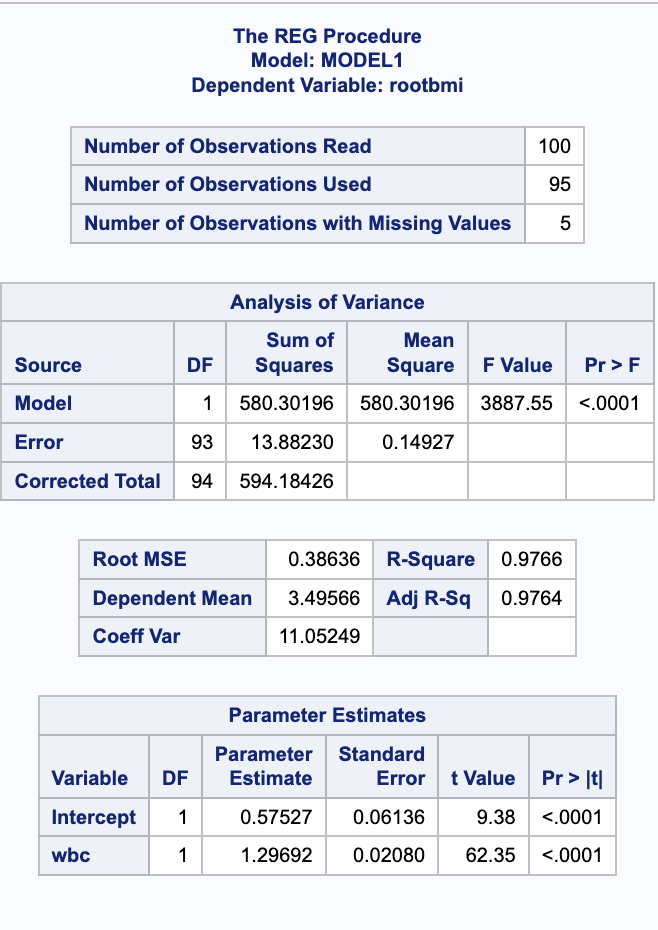
ods graphics on;

proc reg data=Prob8b;

model rootbmi=wbc;

run;

OUTPUT:



Chart

Description automatically generated with medium confidence

By adjusting our data using the **square root transformation**, we can see that our R2 value is immediately improved. This transformation now shows a good fit. Below is a grid of the diagnostics:

Chart

Description automatically generated

The assumptions of constant variance is still shown to be poor assumptions through their respective diagnostic plot. There is indication of improvement in linearity (QQ-plot) and less influential outliers (Cooks D). This transformation gives us a better, more honest depiction of what the data is showing us. As a result, we have we can determine that white-blood cell count is related/indicative of bmi.

**Q:** How do we interpret after transforming the data? Is it as simple as:

White-blood cell count is *linearly* indicative, *after a root transformation,* of bmi. (?)

And what conclusions can we make even with a lack of constant variance?