Data analysis project # 6. Stat 200H

Load the data set Grades.mtw from the “Data”. This contains SAT “Verbal” and “Math” scores, plus “GPA” information collected on a sample of n=200 college students.

Using Calc > Calculator, create the new variable “min” as RMIN(‘Verbal’,’Math’). The reason for doing this is that this combination variable turns out to be a much better predictor of GPA than any of the two scores individually. We consider the regression of Y=GPA on X=min.

a. Start by creating a scatter plot (Graph > Plot) with a lowess smooth superimposed, and a Regression line plot (Stat > Regression > Fitted line plot). Do these lend support to the use of a regression line to model the systematic component of the relationship between GPA and min? Paste the plots in your report and comment on the smooth, the values of R-sq and s, and the relationship as captured by the regression line.

The smooth shows that mean relationship between GPA and min is well approximated by a line. However there is a lot of variability about this line; in fact, the estimate of the error standard deviation, s ~ 0.54 is quite large (GPA ranges between 0 and 4), and the share of the variability in GPA explained by the regression on min is very low, only ~13%. The fitted regression line has positive inclination (slope b1 ~ 0.003), showing a positive relationship between mean GPA and min. In particular, an increase of 1 point in the min between verbal and math SAT score is associated with an average increase of 0.003 GPA points.
b. Now use Stat > Regression > Regression. Specify GPA as response, and min in the “Predictors” box.

- In “Results”, select the second button from the top. This displays an intermediate length output.
- In “Storage”, select residuals and fits. This creates two new columns RESI1 and FITS1 with residuals and fits from the regression.
- In “Option”, in the box for “Prediction Intervals” write the number 650. For “confidence level” right below, leave the default 95%. This produces in output a CONFIDENCE INTERVAL (level 95%) for the average GPA at min=650, plus a PREDICTION INTERVAL (level again 95%) for GPA at min=650.

Paste the output in your report, and create and paste in also the residual plot with smooth superimposed.

Regression Analysis: GPA versus min
The regression equation is
GPA = 0.807 + 0.00313 min

Predictor Coef SE Coef T P
Constant 0.8067 0.3385 2.38 0.018
min 0.0031264 0.0005767 5.42 0.000

S = 0.5429 R-Sq = 12.9% R-Sq(adj) = 12.5%

Analysis of Variance
Source DF SS MS F P
Regression 1 8.6617 8.6617 29.39 0.000
Residual Error 198 58.3583 0.2947
Total 199 67.0200

Predicted Values for New Observations
New Obs Fit SE Fit 95.0% CI 95.0% PI
1 2.8389 0.0544 ( 2.7316, 2.9462) ( 1.7629, 3.9149)

Values of Predictors for New Observations
New Obs min
1 650
Inference on the slope:

The (population) slope $\beta_1$ is estimated by $b_1 \sim 0.003$. In the output you find a standard error for this estimate. FYI, the formula for it is

$$se(b_1) = \frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

i. Construct a 95% confidence interval for $\beta_1$ using $b_1$ and $se(b_1)$. Recall that assuming normality of the deviations from the population regression line, the multiplier can be found under a t with n-2 = 198 degrees of freedom.

**Inverse Cumulative Distribution Function**

<table>
<thead>
<tr>
<th>Student's t distribution with 198 DF</th>
<th>P( X &lt;= x )</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0250</td>
<td>-1.9720</td>
</tr>
</tbody>
</table>

So the 95% CI is given by $0.003 +/- 1.97 \times 0.0006$.

ii. The output also provides the value of a t test statistic, and the corresponding p-value for testing $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$ (two-sided). Do we have strong enough evidence to reject the null? Explain, and also explain what the null hypothesis (a regression line with slope = 0) means.

The p-value for this test is 0.000 (0 to the third decimal), so we have very strong evidence against the hypothesis that the (population) slope is 0. A slope = 0 would mean that the regression line is flat, i.e. that there is no linear dependence between the average GPA and min. This is what we are rejecting based on the data.

Confidence Interval for mean GPA and Prediction interval for GPA at min=650:

An estimate of the (population) average GPA at min=650, as well as a point prediction of GPA at min=650, is provided by the fitted value i.e. the value read on the estimated regression line in correspondence to min=650. You can find the value of this fit, as well as its standard error as an estimate of the average GPA, in the output. Right next, as you requested, the output gives you 95% confidence interval and prediction interval. Multipliers are again obtained from a t with 198 degrees of freedom.

i. Verify that the 95% confidence interval reported in the output is $\hat{y}(650) +/- 1.97se(\hat{y}(650))$, and interpret it.

$2.8389 +/- 1.97 \times 0.0544 \sim (2.7317, 2.9462)$. With 95% confidence, we can say that the mean GPA for a student whose min between verbal and math SAT score is 650 falls between 2.73 and 2.94.
ii. The 95% prediction interval is expressed by a similar formula:
\[ \hat{y}(650) - 1.97se_{pred}(\hat{y}(650)) \leq y_{pred} \leq \hat{y}(650) + 1.97se_{pred}(\hat{y}(650)) \]. Is it broader or narrower than the 95% confidence interval? Since the fitted value and multiplier are the same, what does this mean?

(1.7629, 3.9149) is much broader than (2.7317, 2.9462). This must be because the prediction standard error of y-hat is larger than the standard error of y-hat as an estimate of the mean GPA.

Residuals:

i. Does the residual plot support the use of a regression line to model the mean relationship between GPA and min? Is there anything noticeable about this residual plot?

Yes it does support the use of a regression line, there is no structure to the mean pattern of these residuals (see lowess smooth). One noticeable thing is that we have three “outliers” (data points with very large negative residuals). These could also be seen in the original scatter plot. We may want to repeat the analysis after removing these points.

ii. Create and paste in the report a normal probability plot for the residuals (Graph > Probability plot, you want to use the RESI1 column you stored). If we treat residuals as the sample proxy for deviations from the (population) regression line, what does this plot tell us about the normality assumption we have implicitly employed above for inference on the slope, and confidence and prediction intervals? Explain.

there is some evidence of heavy tails, but the residuals are consistent with a bell-shaped distribution for the deviation from the regression line. So employing t distributions in our inferences should be ok.
More on confidence and prediction intervals:

Use again Stat > Regression > Fitted line plot. In “Options” check to display confidence bands and prediction bands – leave the confidence level at the default 95%. Paste the resulting plot in your report, and read carefully the following addendum (you only need to paste the plot in the report).

The formulae for what you see in this plot are as follows:

regression line: \( \hat{y}(x) = b_0 + b_1x \)

CI bands: \( \hat{y}(x) + /- 1.97se(\hat{y}(x)) \)

PI bands: \( \hat{y}(x) + /- 1.97se_{pred}(\hat{y}(x)) \)

with

\[
se(\hat{y}(x)) = s \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}
\]

\[
se_{pred}(\hat{y}(x)) = \sqrt{s^2 + se^2(\hat{y}(x))}
\]

Things to notice:

- The standard error of y-hat as an estimate of the average GPA for a given x, becomes larger as the given x moves away from x-bar, the center of the x-data (here x-bar is 583.18). Thus, the 95% CI becomes broader as we move from the center towards the periphery of the data cloud, expressing the fact that our uncertainty in estimating the average GPA increases. See how the band broadens.

- The prediction standard error of y-hat is larger than the standard error of y-hat as an estimate of the average GPA. See how the prediction band exceeds the CI band. Here is the rationale: if we want to estimate the average response at a given x, uncertainty is created by the fact that we are estimating the regression line on our data. This is captured by se of y-hat. However, if we want to predict a new response observation at
that $x$, we also need to take into account the uncertainty created by the deviations from the (population) regression line. This is captured by $s$ (our estimate of the standard error of these deviations). In the formula for the prediction standard error of $y$-hat, we are adding two variances, and then taking the square root. The reason for adding the two variances is as follows: we assume the new observation we are trying to predict (hence, the deviation that will come with it) to be independent from the observations we had in our sample (hence, the regression line estimation and the fitted value).