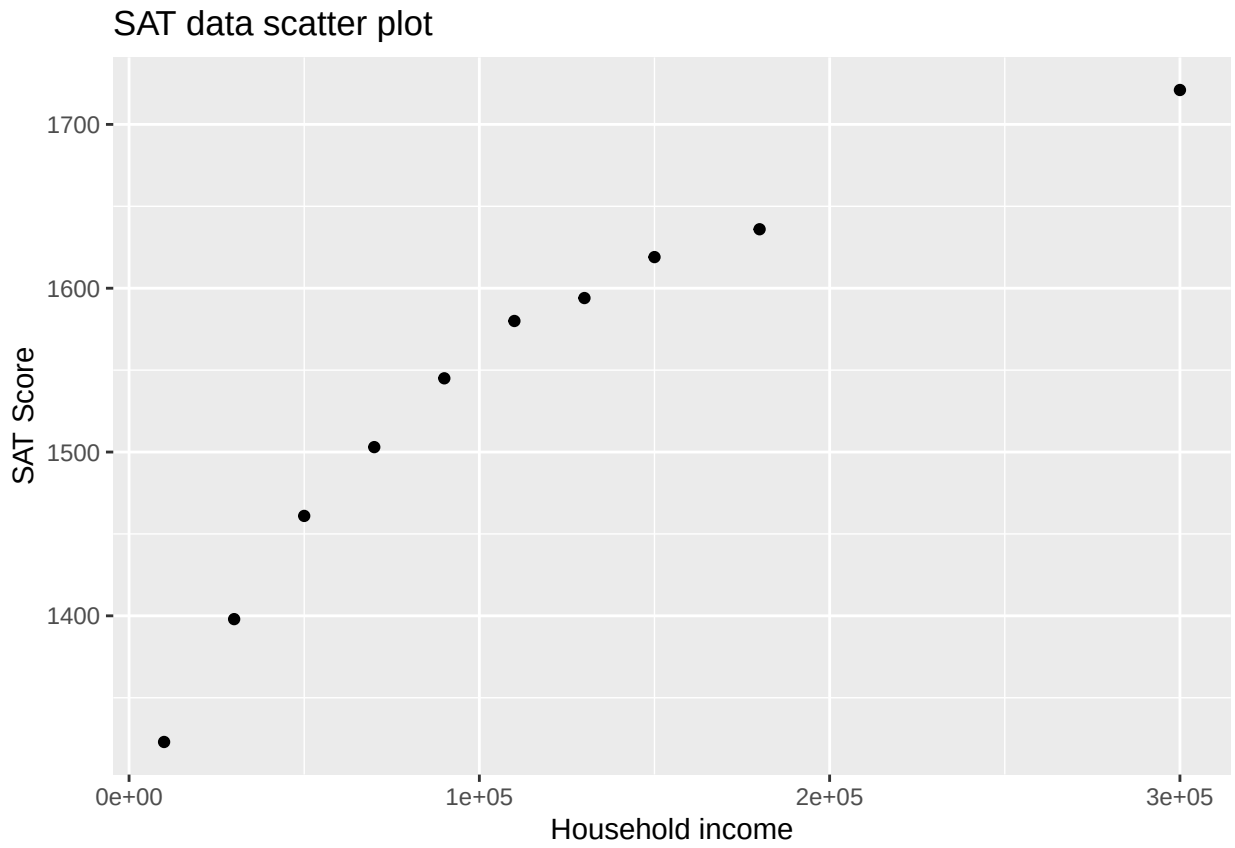
Stat 21

Class 14 Worksheet Solutions

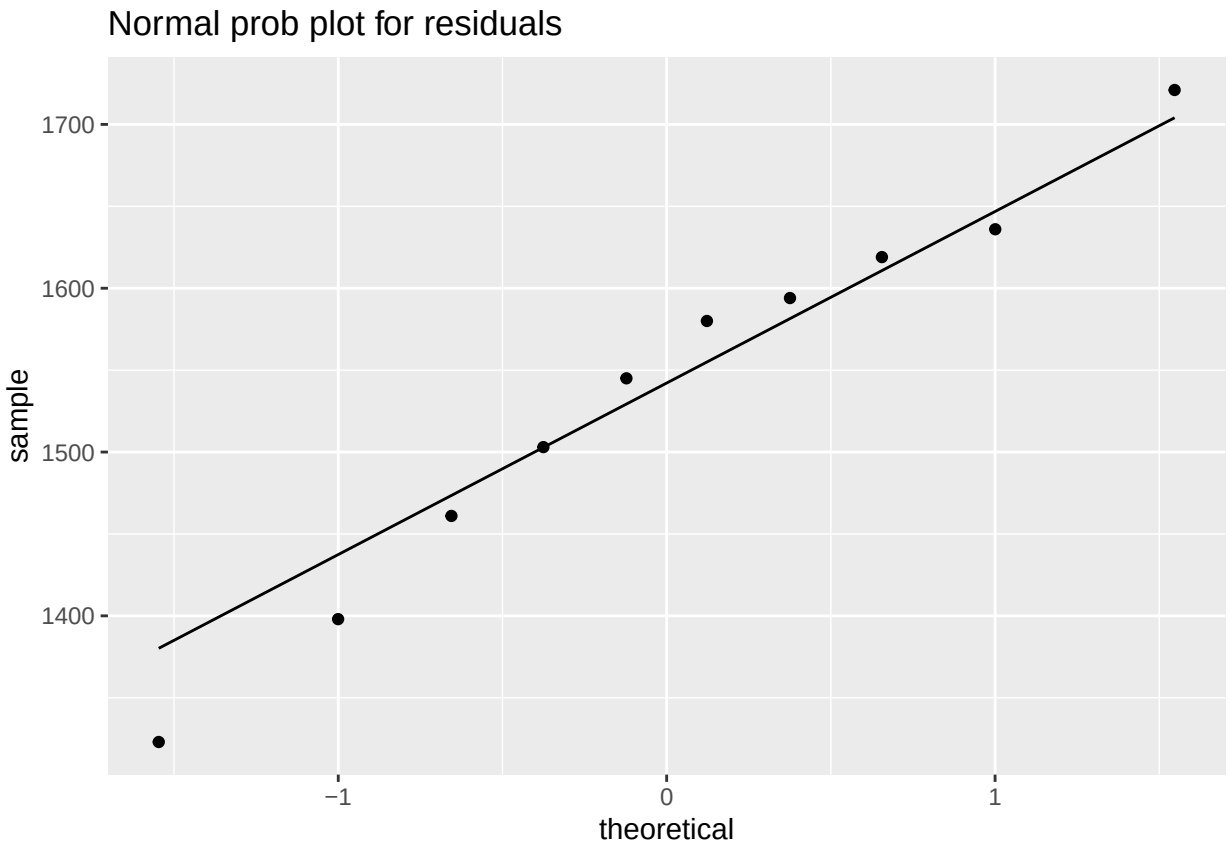
```
income_SAT2 <- tibble(income = c(10000, 30000, 50000, 70000, 90000, 110000,
                                 130000, 150000, 180000, 300000),
                      SAT_score = c(1323, 1398, 1461, 1503, 1545, 1580, 1594,
                                    1619, 1636, 1721))
SLR_SAT <- lm(SAT_score ~ income, income_SAT2)
summary(SLR_SAT)
```

```
##
## Call:
## lm(formula = SAT_score ~ income, data = income_SAT2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -82.21 -24.00  14.58  32.31  44.60
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.392e+03  2.531e+01  55.009 1.32e-11 ***
## income      1.302e-03  1.834e-04   7.098 0.000102 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.75 on 8 degrees of freedom
## Multiple R-squared:  0.863,  Adjusted R-squared:  0.8458
## F-statistic: 50.38 on 1 and 8 DF,  p-value: 0.0001022
```

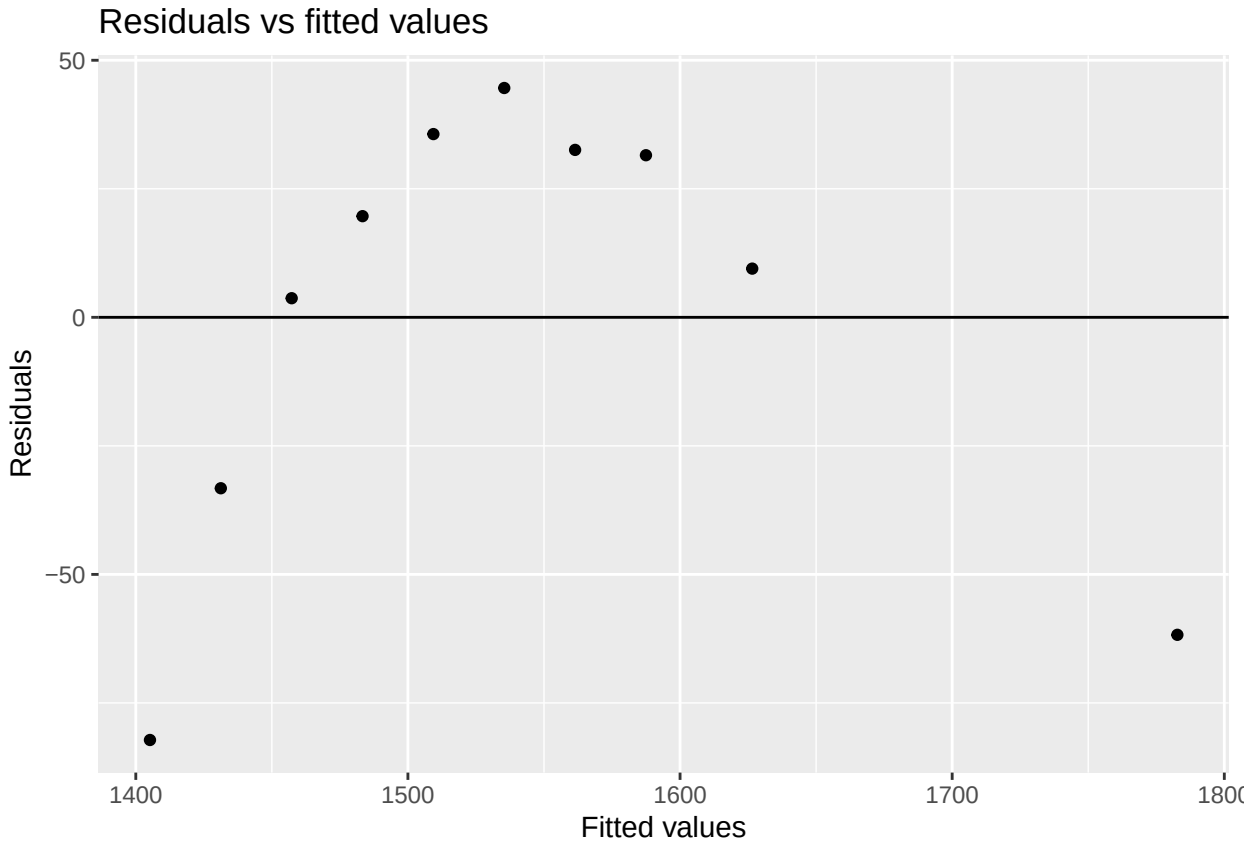
```
ggplot(income_SAT2, aes(x=income ,y=SAT_score)) +
  geom_point() +
  labs(title="SAT data scatter plot", x="Household income", y="SAT Score")
```




```
income_SAT2 <- income_SAT2 %>% mutate(resids = SLR_SAT$residuals,
                                      fits = SLR_SAT$fitted.values)
ggplot(income_SAT2, aes(sample=SAT_score)) +
  stat_qq() +
  stat_qq_line() +
  labs(title="Normal prob plot for residuals")
```




```
ggplot(income_SAT2, aes(x=fits, y=resids)) +
  geom_point() +
  labs(title="Residuals vs fitted values", x="Fitted values", y="Residuals") +
  geom_hline(yintercept=0)
```



Step 1

Regression equation:

$$\text{Average SAT Score} = 1392 + 0.001302(\text{Household Income})$$

Interpretation of slope:

For every dollar increase in household income, the average SAT score increases by 0.0013 points. In other words, for every \$10,000 increase in income, we expect SAT scores to increase by 13 points, on average.

Comments on plots:

There is a strong, non-linear relationship evident in the residual plot. This indicates that the relationship between the predictor and response is **not linear**. Furthermore, it is difficult to determine if the constant variance assumption is met because of this pattern.

Step 2

```
income_SAT_trans <- income_SAT2 %>% mutate(SAT_score_trans = SAT_score^10)
SLR_SAT_trans <- lm(SAT_score_trans ~ income, data = income_SAT_trans)
summary(SLR_SAT_trans)
```

```
##
## Call:
## lm(formula = SAT_score_trans ~ income, data = income_SAT_trans)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -4.292e+30 -2.056e+30 -1.117e+30  1.658e+30  6.678e+30
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.557e+30  1.963e+30   4.869  0.00124 **
## income      7.338e+26  1.422e+25  51.589 2.21e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.625e+30 on 8 degrees of freedom
## Multiple R-squared:  0.997,  Adjusted R-squared:  0.9966
## F-statistic:  2661 on 1 and 8 DF,  p-value: 2.208e-11
```

Regression equation for transformed data:

$$\text{Average (SAT Score)}^{10} = 9.557 \times 10^{30} + (7.338 \times 10^{26})(\text{Household Income})$$

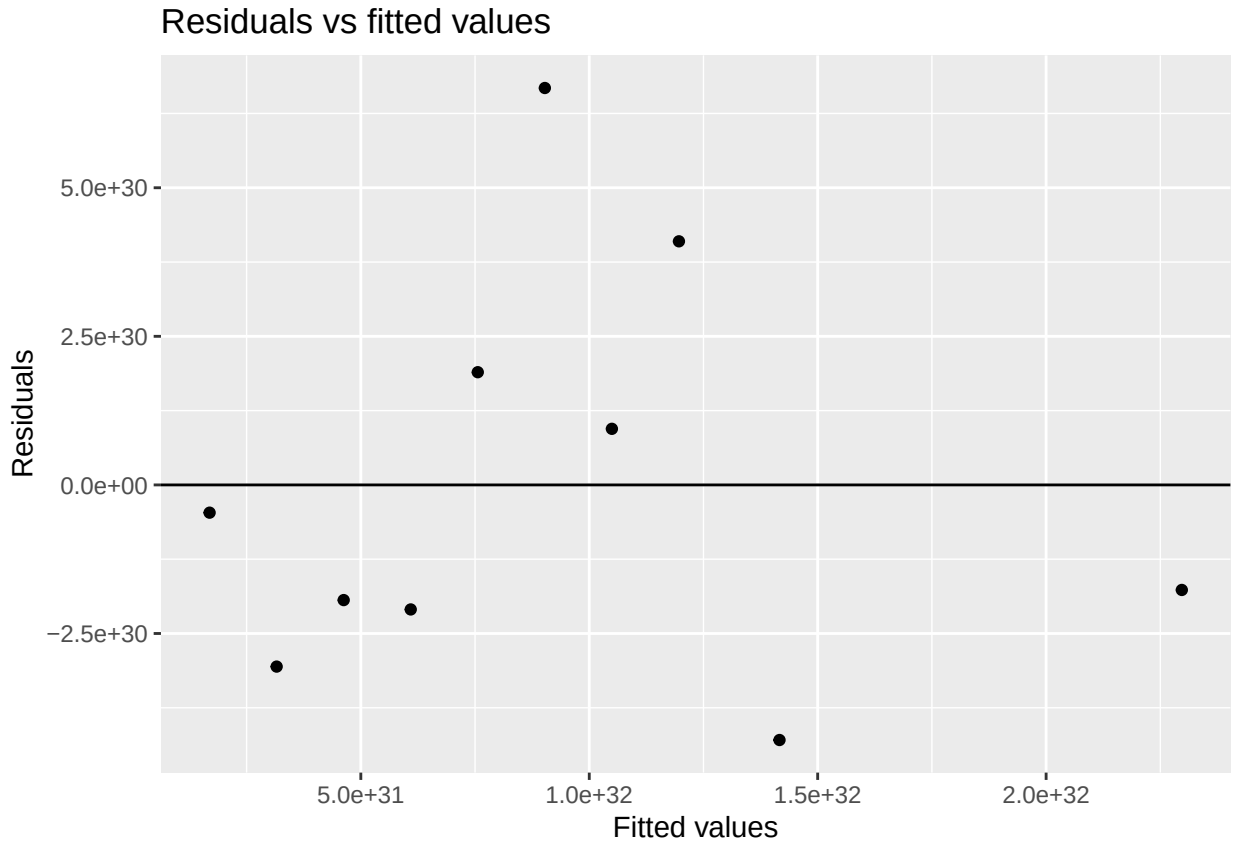
Interpretation of the slope:

For every dollar increase in household income, we expect the SAT score raised to the 10th power to increase by 7.338×10^{26} points, on average.

In the units of the original response variable, this means that for every increase in the logarithm of household income (base 10) (i.e. for each increase in log-dollars), we expect the SAT score to increase by $log_{10}(7.338 \times 10^{26}) = 26.87$ points.

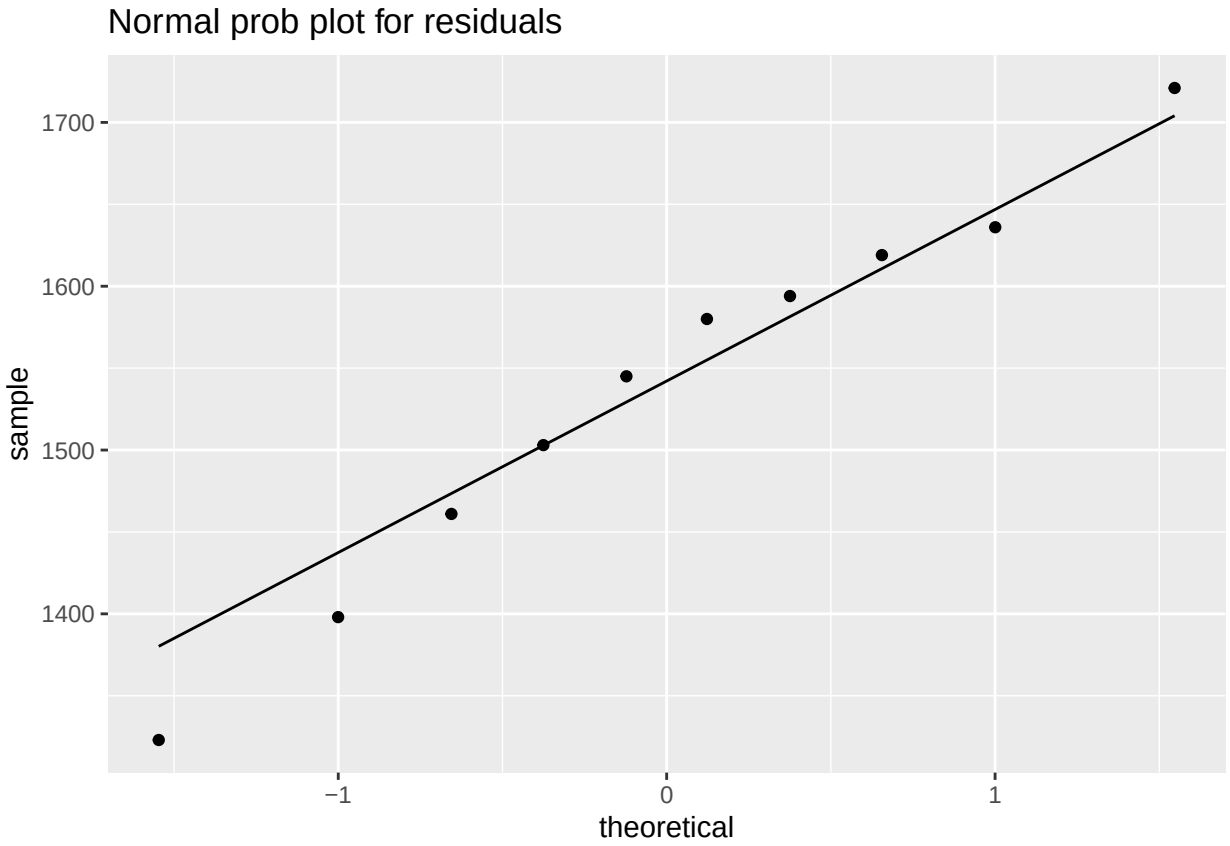
Residual Plot

```
income_SAT_trans2 <- income_SAT_trans %>% mutate(resids = SLR_SAT_trans$residuals,
                                                 fits = SLR_SAT_trans$fitted.values)
ggplot(income_SAT_trans2, aes(x=fits, y=resids)) +
  geom_point() +
  labs(title="Residuals vs fitted values", x="Fitted values", y="Residuals") +
  geom_hline(yintercept=0)
```



Normal Probability Plot

```
ggplot(income_SAT_trans2, aes(sample=SAT_score)) +
  stat_qq() +
  stat_qq_line() +
  labs(title="Normal prob plot for residuals")
```



Comments on plots:

The residuals plot looks much better in terms of not showing any particular pattern (as we want). There doesn't seem to be any strong change in the spread of the residuals about the line at $e = 0$ for different values of $\hat{y}$ so the constant variance assumption seems reasonable. The sample quantiles of the residuals match pretty well with the theoretical quantiles of a normal distribution so it also looks like the assumption that the errors are normally distributed is reasonable as well.

Step 3