ST 503 Homework 4

Eric Warren

2024-02-10

Contents

1 Problem 1

Consider a linear model with only intercept: $y_i = \mu + e_i$, i = 1, . . . , n and $e_1, ..., e_n$ are IID $N(0, \sigma^2)$.

1.1 Part A

Find the $100(1-\alpha)\%$ confidence interval for μ.

First let us note that we know the following is true:

- $P(c^T\hat{\beta} - t_{\frac{\alpha}{2},n-rank(X)}\sqrt{c^T\hat{\sigma^2}(X^TX)^-c} \leq c^T\beta \leq c^T\hat{\beta} + t_{\frac{\alpha}{2},n-rank(X)}\sqrt{c^T\hat{\sigma^2}(X^TX)^-c}) = 1-\alpha$
- The $100(1-\alpha)\%$ confidence interval is then $c^T\hat{\beta} \pm t_{\frac{\alpha}{2},n-rank(X)}\sqrt{c^T\hat{\sigma^2}(X^TX)^-c}$

Knowing this, we can find what c, c^T, X, X^T are to get our generic formula for a model in which it is $y_i = \mu + e_i$. First note that in this case $X\beta = \mu$. In this case, we know that β is a matrix of the predictors which in this case is just $\beta = \begin{pmatrix}\mu\end{pmatrix}$ since this is only predictor. We also know that in order for each y_i to get μ as its predictor term, we need a n x 1 matrix representing X where all the values are 1 to get this to equal μ. Thus, we can say that $X = \begin{pmatrix} 1 \\ 1 \\ ...(1) \\ 1 \end{pmatrix}$. Now knowing this, we know that since X is a n x 1 matrix we know that X^T is a 1 x n matrix (or can say vector) such that $X^T = \begin{pmatrix}1 & 1 & ...(1) & 1\end{pmatrix}$. Now, this will help us determine that $X^T X = \begin{pmatrix}1 & 1 & ...(1) & 1\end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ ...(1) \\ 1 \end{pmatrix} = n$. We will use this solution that $X^T X = n$ for later. Lastly, we can clearly see the rank(X) = 1 since there is only one (independent) column present in our design matrix X.

Now here we are trying to estimate our parameters in this case it is just estimating μ with using $\hat{\mu}$. So we know we need $c^T\hat{\beta} = \hat{\mu}$. First, we know that $\beta = \begin{pmatrix}\mu\end{pmatrix}$ so $\hat{\beta} = \begin{pmatrix}\hat{\mu}\end{pmatrix}$. So now to plug in that $c^T\hat{\beta} = c^T(\hat{\mu}) = \hat{\mu}$ we know that $c^T = 1$ which in return also means that $c = 1$.

Now let us plug things in knowing that the $100(1-\alpha)\%$ confidence interval is $c^T\hat{\beta} \pm t_{\frac{\alpha}{2}, n-rank(X)}\sqrt{c^T\hat{\sigma^2}(X^TX)^-c}$. We can say that:

$$
\begin{aligned}
& c^T\hat{\beta} \pm t_{\frac{\alpha}{2}, n-rank(X)}\sqrt{c^T\hat{\sigma^2}(X^TX)^-c} \\
&= 1\hat{\mu} \pm t_{\frac{\alpha}{2}, n-1}\sqrt{1\hat{\sigma^2}(n)^-(1)} \\
&= \hat{\mu} \pm t_{\frac{\alpha}{2}, n-1}\sqrt{\hat{\sigma^2}(n)^-} \\
&= \hat{\mu} \pm t_{\frac{\alpha}{2}, n-1}\sqrt{\hat{\sigma^2} * \frac{1}{n}} \\
&= \hat{\mu} \pm t_{\frac{\alpha}{2}, n-1}\sqrt{\frac{\hat{\sigma^2}}{n}}
\end{aligned}
$$

So from this we can see that the $100(1-\alpha)\%$ confidence interval for our intercept only model is $= \hat{\mu} \pm t_{\frac{\alpha}{2}, n-1}\sqrt{\frac{\hat{\sigma^2}}{n}}$. Please note in most statistics formulas or books we see that $\hat{\mu}$ is replaced with $\bar{y}$ since we know that the intercept in our model is going to be the average of our y values, but for purposes of keeping notation easy to follow I left the solution using $\hat{\mu}$.

1.2 Part B

Find the $100(1-\alpha)\%$ prediction interval for a new observation y_{new}.

First let us note that we know the following is true:

- The $100(1-\alpha)\%$ prediction interval is then $X_{new}^T\hat{\beta} \pm t_{\frac{\alpha}{2}, n-rank(X)}\sqrt{X_{new}^T\hat{\sigma^2}(X^TX)^-X_{new} + \hat{\sigma^2}}$

Knowing this, we can find what $X_{new}, X_{new}^T, X, X^T$ are to get our generic formula for a model in which it is $y_i = \mu + e_i$. First note that in this case $X_{new}\hat{\beta} = \hat{\mu}$. In this case, we know that $\hat{\beta}$ is a matrix of the predictors which in this case is just $\hat{\beta} = \begin{pmatrix}\hat{\mu}\end{pmatrix}$ since this is only predictor. We also know that in order for each y_i to get μ as its predictor term, we need a 1 x 1 matrix representing X_new where its value being 1 to get this to equal $\hat{\mu}$. Thus, we can say that $X_{new} = \begin{pmatrix}1\end{pmatrix}$. Now knowing this, we know that since X_{new} is

a 1 x 1 matrix we know that X_{new}^T is a 1 x 1 matrix such that $X_{new}^T = \begin{pmatrix}1\end{pmatrix}$. Now also note from **Part A**, we already know that $X^TX = \begin{pmatrix}1 & 1 & ...(1) & 1\end{pmatrix}\begin{pmatrix}1\\1\\...(1)\\1\end{pmatrix} = n$. We will use this solution that $X^TX = n$ for later. Lastly, we can clearly see the rank(X) = 1 since there is only one (independent) column present in our design matrix X.

Now let us plug things in knowing that the $100(1-\alpha)\%$ prediction interval we should get is $X_{new}^T\hat{\beta} \pm t_{\frac{\alpha}{2},n-rank(X)}\sqrt{X_{new}^T\hat{\sigma^2}(X^TX)^-X_{new} + \hat{\sigma^2}}$. We can say that:

$$X_{new}^T\hat{\beta} \pm t_{\frac{\alpha}{2},n-rank(X)}\sqrt{X_{new}^T\hat{\sigma^2}(X^TX)^-X_{new} + \hat{\sigma^2}}$$

$$1\hat{\mu} \pm t_{\frac{\alpha}{2},n-1}\sqrt{1 * \hat{\sigma^2}(n)^- * 1 + \hat{\sigma^2}}$$

$$\hat{\mu} \pm t_{\frac{\alpha}{2},n-1}\sqrt{\hat{\sigma^2}\frac{1}{n} + \hat{\sigma^2}}$$

$$\hat{\mu} \pm t_{\frac{\alpha}{2},n-1}\sqrt{\hat{\sigma^2}(\frac{1}{n} + 1)}$$

So from this we can see that the $100(1-\alpha)\%$ prediction interval for our intercept only model is $\hat{\mu} \pm t_{\frac{\alpha}{2},n-1}\sqrt{\hat{\sigma^2}(\frac{1}{n}+1)}$. Please note in most statistics formulas or books we see that $\hat{\mu}$ is replaced with $\bar{y}$ since we know that the intercept in our model is going to be the average of our y values, but for purposes of keeping notation easy to follow I left the solution using $\hat{\mu}$.

1.3 Part C

Find the leverage values of all the observations.

We know that all of the leverage values should sum up to rank(X). Note how we said before that rank(X) = 1, so $\sum h_{ii} = rank(X) = 1$. Now also note since $\sum h_{ii} = rank(X)$ the average of the leverage or $\bar{h_{ii}} = \frac{1}{n}\sum h_{ii} = \frac{rank(X)}{n} = \frac{1}{n}$ (which was calculated by just multiplying both sides by $\frac{1}{n}$ for the equation $\sum h_{ii} = rank(X) = 1$). Now we also know that all the values of x_i are the same from knowing that our matrix $X = \begin{pmatrix}1\\1\\...(1)\\1\end{pmatrix}$. Thus, we can say that if all x_i are the same then all $x_i = \bar{x}$. Since this is the case and knowing that our leverage $h_{ii} = \frac{1}{n}$ when $x_i = \bar{x}$ or knowing that since all x_i are the same they should have the same leverage (because in an intercept model all our $x_i = 1$), we have shown that for all observations our leverage $h_{ii} = \frac{rank(X)}{n} = \frac{1}{n}$.

2 Problem 2

Using the `sat` dataset, fit a model with the total SAT score as the response and expend, salary, ratio and takers as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant. Do not provide any plots about which you have nothing to say.

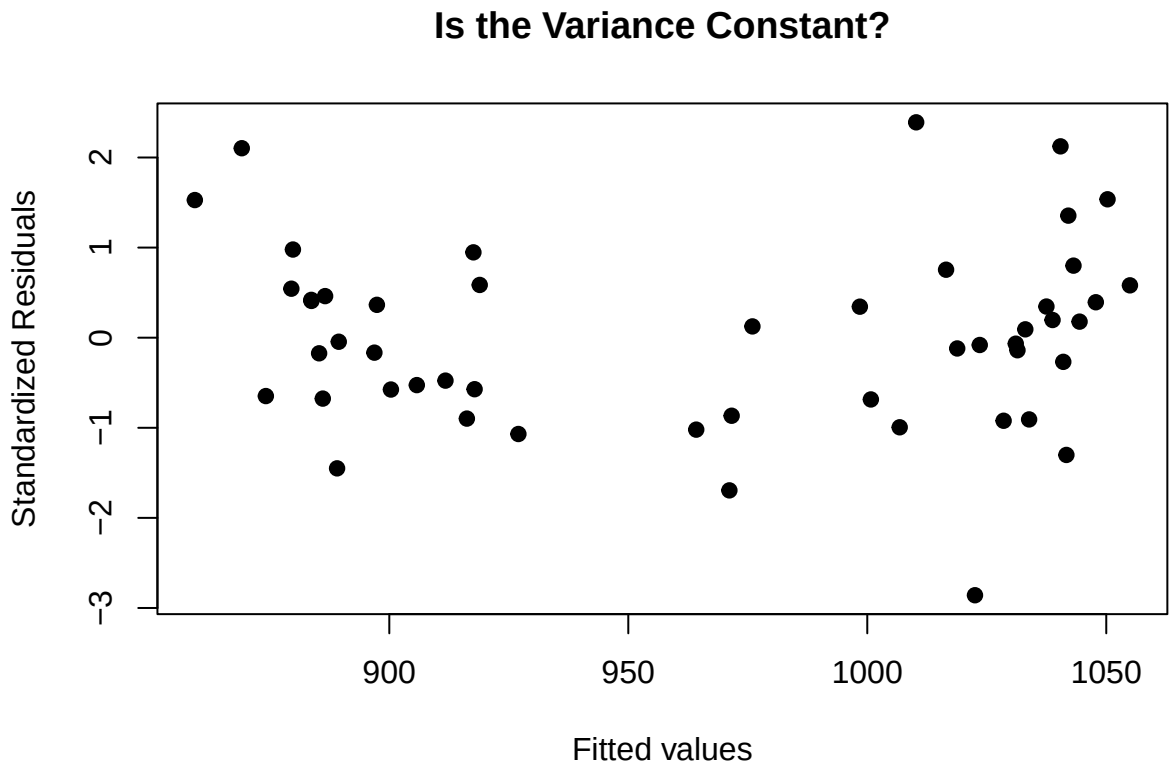
```
library(faraway)
model <- lm(total ~ expend + salary + ratio + takers, sat)
```

2.1 Part A

Check the constant variance assumption for the errors.

We are going to make a plot here and take a look to see if there are any patterns in the residuals as compared with the fitted values. If not, then we can say there is constant variance.

```
plot(fitted(model),
     rstandard(model),
     xlab = "Fitted values",
     ylab = "Standardized Residuals",
     main = "Is the Variance Constant?",
     pch = 19)
```



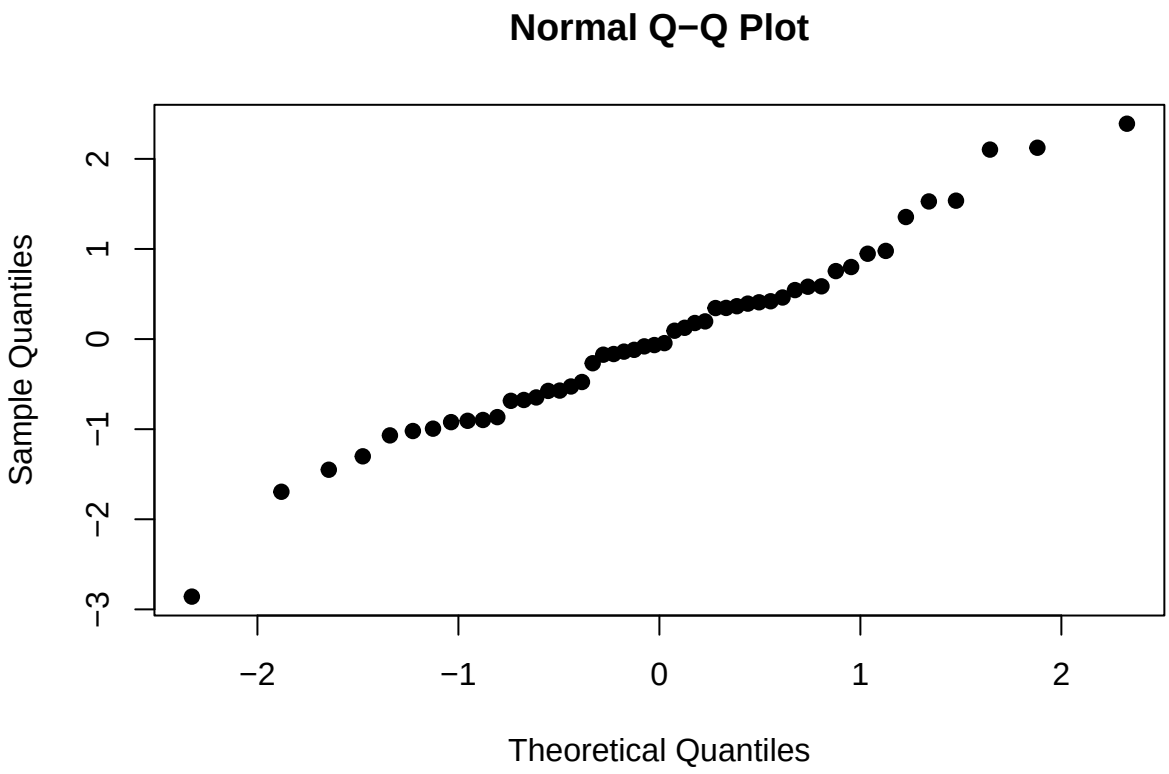
My interpretation of the plot is that there is constant variance since the variance seems to be pretty random with no pattern in place. Since the residuals seem to be random with no patterns present then I would say the constant variance assumption is satisfied.

2.2 Part B

Check the normality assumption.

We can check if it is normal by looking at a Q-Q plot. If the Q-Q plot shows a linear trend (at that 45 degree angle increase) this makes us assume the normality assumption is satisfied.

```
qqnorm(rstandard(model), pch = 19)
```



While it is not perfect it seems to have this linear upward trend that we are looking for. Since it is not fully obvious though we can check with a Shapiro-Wilk normality test. If the resulting p-value is less than 0.05 then we can say the normality assumption is violated.

```
shapiro.test(rstandard(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(model)
## W = 0.98021, p-value = 0.5607
```

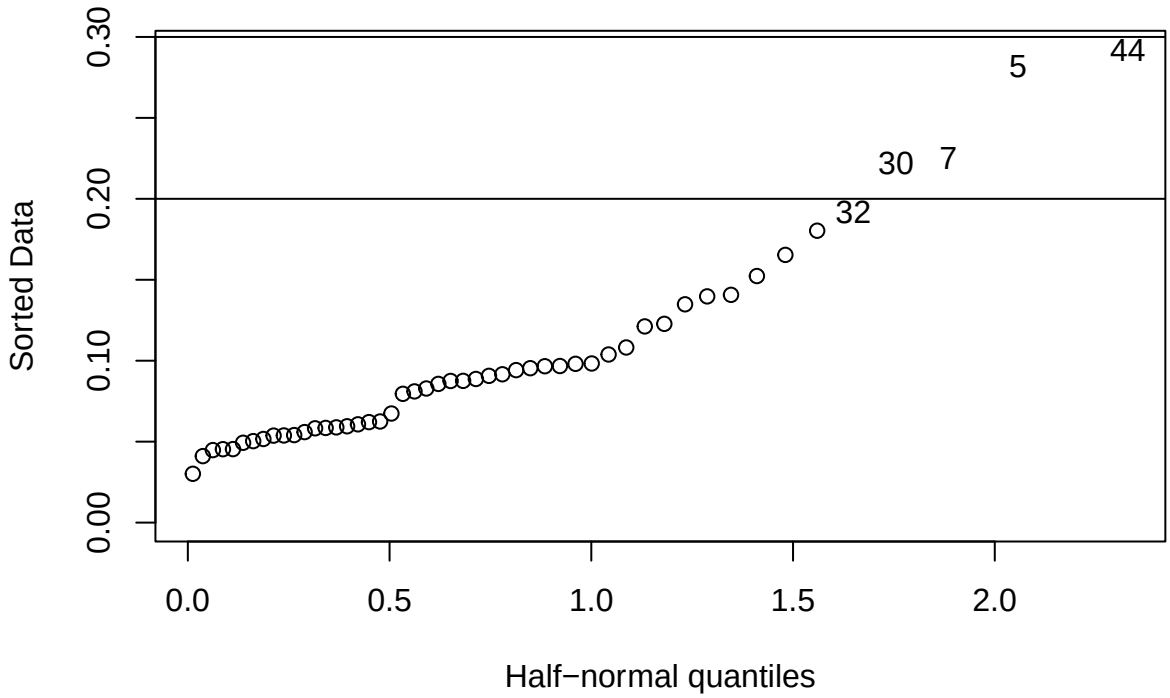
Since the resulting p-value of the test is 0.5607232 we can say that we fail to reject our hypothesis that the normality assumption is satisfied and with a high p-value we have a pretty reasonable suspicion that the normality assumption is satisfied. For this reason, I would say that we can say that our normality assumption is validated by these two things shown.

2.3 Part C

Check for large leverage points.

Here we are going to find which points have high leverage. First let us find the 5 highest leverage points just to give us an idea of what we are looking at.

```r
# get leverage values
lv <- hatvalues(model)

# halfnormal plot, mark 5 points with highest leverage
halfnorm(lv, nlab = 5)

# 3 times average of leverages
abline(h = 3*mean(lv))

# 2 times average of leverages
abline(h = 2*mean(lv))
```



As we can see observations 44, 5, 7, 30, and 32 are the top 5 point with the most leverage. However we see that from our plot that the only points that are considered high leverage points if using 3 times the leverage mean are no points and only 2 times the leverage mean being the first 4 observations shown. We show this below.

```r
# Which points are really high leverage
which(lv >= 3 * mean(lv))
```

```
## named integer(0)
```

```r
# Which points have considered high leverage
which(lv >= 2 * mean(lv))
```

```
##  California Connecticut  New Jersey        Utah
##           5           7          30          44
```

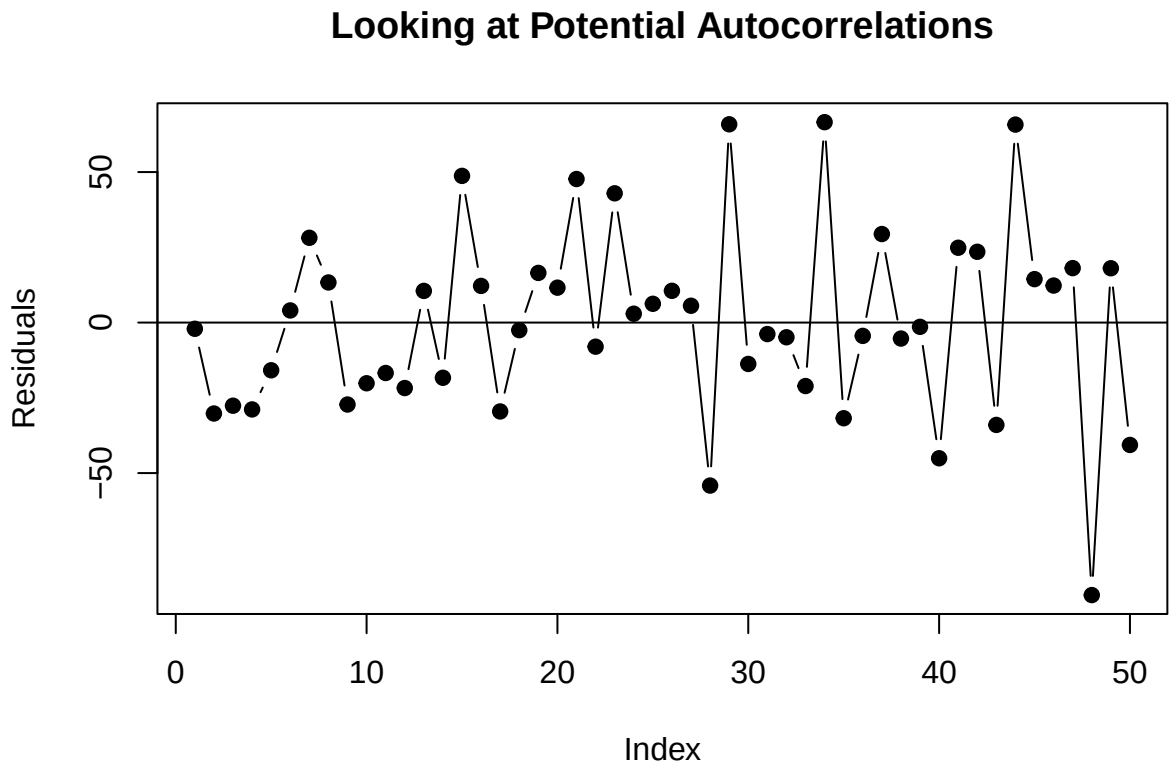
As we can see the points with considered super high leverage are really none, however points 5 and 44 seem to be the highest which are the states of California and Utah for SAT scores. We might also want to double check how things look in Connecticut and New Jersey as well since they have some high-ish levels of leverage. All in all, since no observations have more than 3 times the mean of our leverage values, we really have nothing overly concerning which is good (other than monitoring California and Utah since they are still higher ones).

2.4 Part D

Check for serial correlation in the errors.

Let us make an index plot (plot of residuals against their indices), and test for autocorrelation using the Durbin-Watson test. First let us look at the index plot to see if anything looks strange.

```
plot(resid(model),
     ylab = "Residuals",
     main = "Looking at Potential Autocorrelations",
     pch=19,
     type="b")
abline(h=0)
```



As we can see in this first plot, we do not see any patterns with the residuals making us think autocorrelation between values is most likely not present. We can double check that using the Durbin-Watson Test. If our p-value is low (below 0.05), then autocorrelation is present.

```
library(lmtest)
dwtest(formula = total ~ expend + salary + ratio + takers, data = sat)
```

```
##
##  Durbin-Watson test
##
## data:  total ~ expend + salary + ratio + takers
## DW = 2.4525, p-value = 0.9459
## alternative hypothesis: true autocorrelation is greater than 0
```

With a really high p-value of 0.9459417, we can definitely fail to reject the statement of autocorrelation being present and if anything have a reasonable suspicion thinking it is not present. Due to the index plot above and the high p-value in the Durbin-Watson test, I would say that we do not have autocorrelation in the errors; therefore, I would say the serial correlation in the errors is fine with none (or trace amounts) being present.

3 Problem 3

3.1 Part A

Explain why a prediction interval of a new observation y_{new} is wider than the corresponding confidence interval for $E(y_{new})$ in a linear model as discussed in class.

We know that from class the prediction interval is wider because we are adding an extra σ^2 term and unlike a confidence interval standard error which can vanish as $n \to \infty$, the prediction interval standard error will start by decreasing and then stabilize. The reason that we add this extra σ^2 to a prediction interval is because we have to account for the variance (or error) we get from the new observation y_{new} itself. That is why the variance formula for a prediction interval $= V(X_{new}^T\hat{\beta}) + V(y_{new}) = X_{new}^T\sigma^2(X^TX)^-X_{new} + \sigma^2 > c^T\sigma^2(X^TX)^-c$ which the latter is that of a confidence interval. All in all, the variance (or error) we get from the new observation y_{new} itself is why a prediction interval is wider than the corresponding confidence interval for $E(y_{new})$ in a linear model as discussed in class.

3.2 Part B

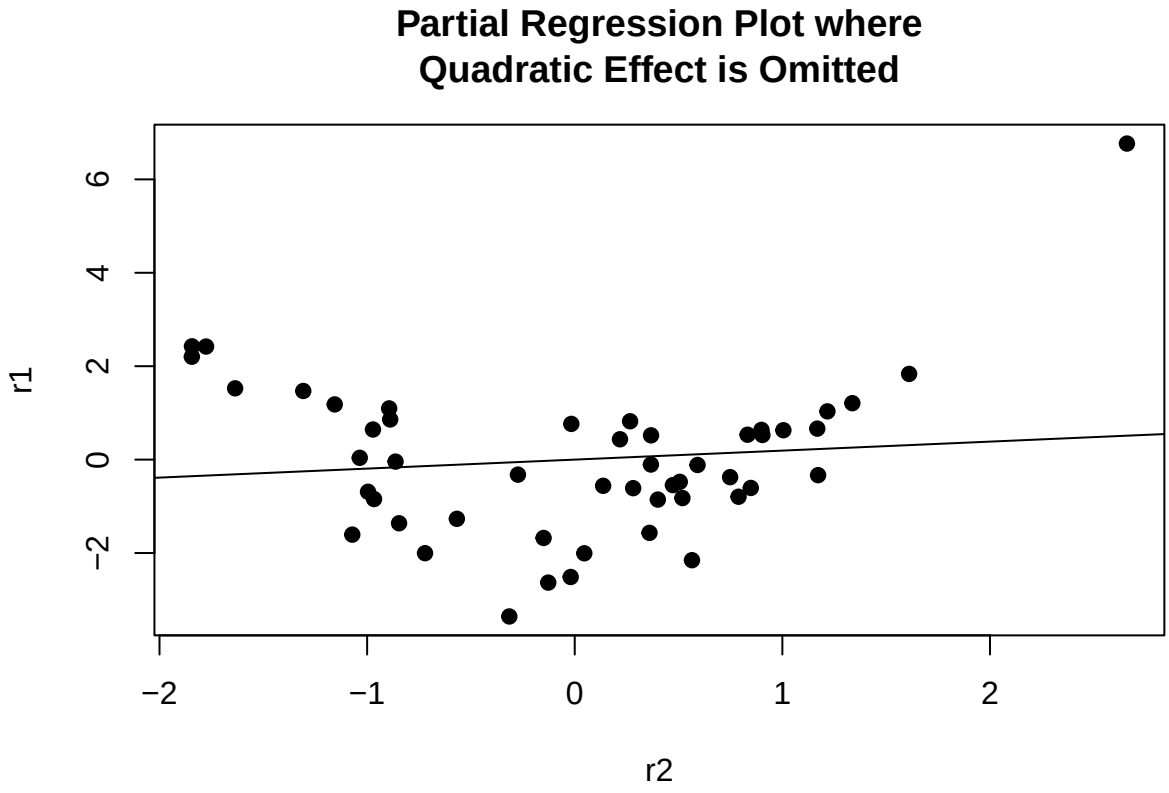
Explain why misspecification of the systematic component, $X\beta$, may lead to nonlinear patterns in the residual plot (residual vs fitted value). Simulate a dataset to demonstrate this point.

The nonlinear trend in the residual plot will come from not having an adequate model. Sometimes we will have this nonlinear trend which signals that we need to update our model to also have a nonlinear term. So by misspecifying our systematic component (and thinking we should only have linear terms as predictors), there is a chance that this hunch is truly wrong and the predictor follows a nonlinear trend which in return will make our residuals also non-linear. Let us take an example using simulated data which shows this more in depth.

First we are going to generate some random data and fit it to a linear terms model despite seeing that x1 should be quadratic.

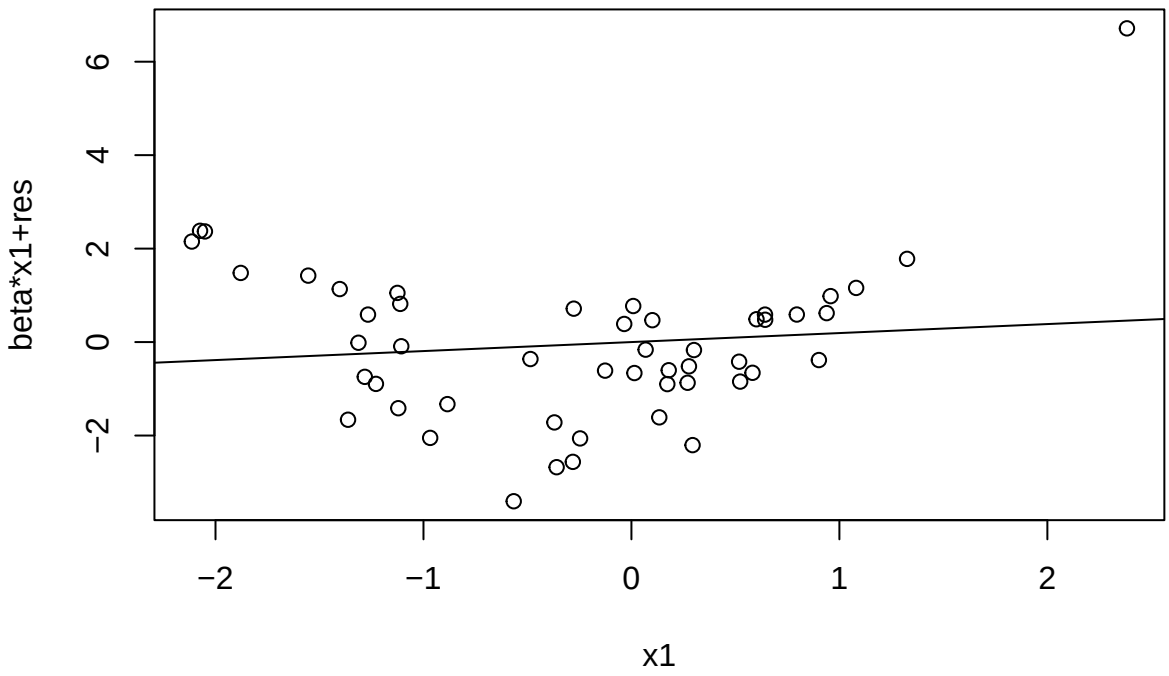
```
# Get random data
set.seed(999)
x1 <- rnorm(50)
x2 <- rnorm(50)
```

```r
y <- 1 + x1 + x1^2 + x2 + rnorm(50)

# Fit a linear term model and compare our coefficients for specification
wrong_lm <- lm(y ~ x1 + x2)
mod1 <- lm(y ~ x2)
mod2 <- lm(x1 ~ x2)
r1 <- resid(mod1)
r2 <- resid(mod2)
pr_lm_sim <- lm(r1 ~ r2)

# Plot this data to show how the residuals look
plot(r2, r1, pch = 19, main = "Partial Regression Plot where
Quadratic Effect is Omitted")
abline(coef(pr_lm_sim))
```



As we can see here the partial regression plot shows that the trend for x1, our first predictor, is not following a linear trend and is really quadratic. We can see with the next plot that the result of this is also going to show residuals where they are also showing a non-linear trend.

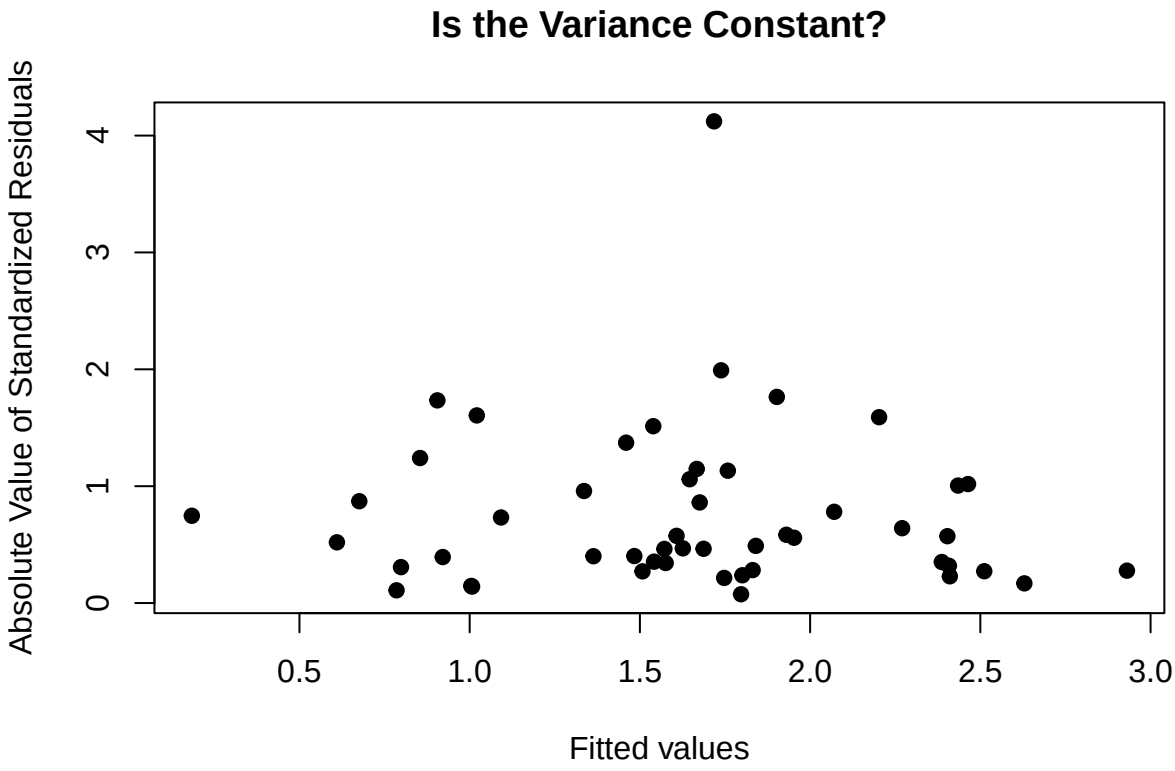
```r
prplot(wrong_lm, 1)
```



In the synthetic data, we can clearly see a nonlinear trend in our partial residual plot indicating the model needs to be updated with a nonlinear term.

Lastly, our actual residual plot for the whole model we created shows this non-constant.

```
plot(fitted(wrong_lm),
     abs(rstandard(wrong_lm)),
     xlab = "Fitted values",
     ylab = "Absolute Value of Standardized Residuals",
     main = "Is the Variance Constant?",
     pch =19)
```



As we can see the absolute residual value peaks towards the middle of our fitted values and then decreases towards the higher values showing this quadratic like trend (non-linear) in our fitted vs residual plot. So because we misspecified our x1 variable this resulted in non-linear trends in our residual plots.