PSTAT 126: Homework #3 Solutions

Problem 1: Obtaining data samples from the dataset "HW 3 Dataset 1" (which will be sent in an email with that subject line), utilize any or all of the various regression model-fit assessment visual plots and/or graphs, results of statistical hypothesis tests, and numerical regression model accuracy measures, as well as any other applicable diagnostic tools available in R that we have discussed in the course thus far, to help identify and validate a linear regression model that appears to best fit the given data. These will likely include for example direct X-Y variable graphs and residuals vs. fitted value plots implemented in R. You will likely need to try a number of different response and/or predictor variable transformations, for example, using the diagnostic tools to decide which regression models appear consistent with the data and which do not. You should identify one or perhaps two candidate regression models that you believe are most consistent with the given data. Please try to include screenshots of the graphical plots you use as well as quote any relevant R code output. You can or even should include the R code itself as well if you feel it helps to support your argument. Please explain your reasoning as to why the model(s) you propose may be the right one(s) in plain natural language prose.

Solution: The R code used to generate the data for this problem is the following:

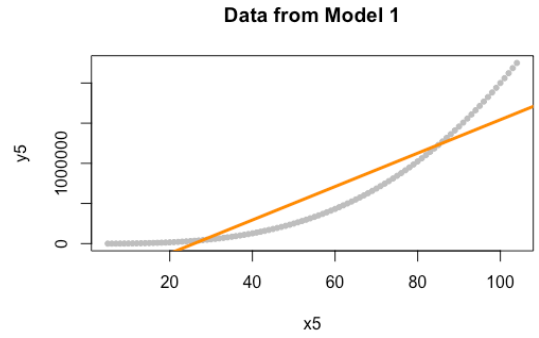
```
x5 <- seq(5,104)
e5 <- rnorm(100, mean = 0,  sd = 0.5)
y5 <- 3 + 2*(x5)*(x5)*(x5) + e5

data_file <- data.frame(y5,x5)
```

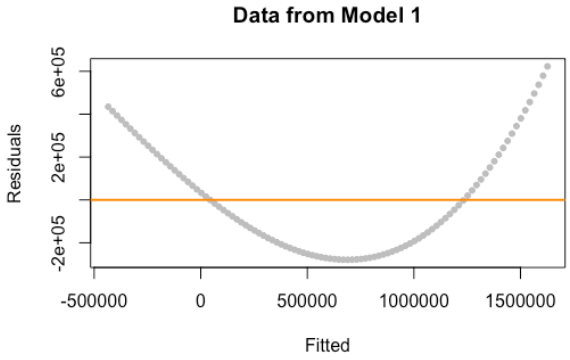
This corresponds to the polynomial regression model

$$Y_n = 3 + 2x_n^3 + \epsilon_n,\ n = 1, \dots, N, \qquad \text{(Model 1)}$$
$$\epsilon_n \sim N(0, 0.25), n = 1, \dots, N,$$

with N=100.

Of course, given that you did not generate the data yourself, you would not know this and would likely have to experiment with different candidate regression models to see which one might offer the best fit. Note that the direct Predictor vs. Response (X-Y variable) plot of the generated data along a regression line according to a Simple Linear Regression model looks like the graph on the left:





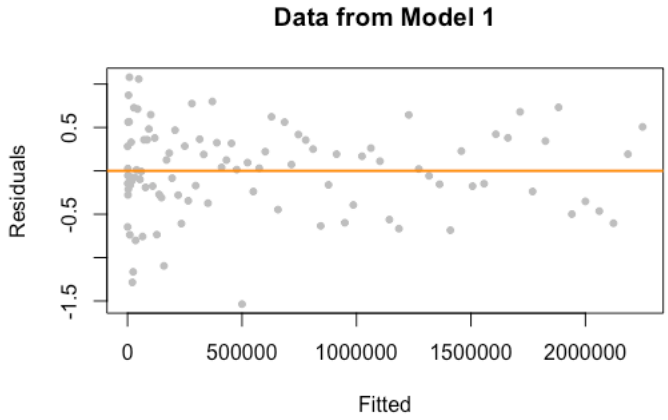
The SLR model clearly does not look like a good fit. The residuals vs. fitted values plot for SLR (on the right above) looks significantly worse and appears to confirm a poor fit. Ideally, there should be a relatively even and symmetric distribution of the data samples across and along the residuals' 0-line (in orange), which is clearly not what we see, implying that the model, in particular our proposed regression (mean) function model, is not capturing much or most of the joint, deterministic (as opposed to probabilistic or stochastic) behavior of the variables. So we need to try a different candidate regression model.

We might decide to try instead a model of the form

$$Y_n=\beta_0+\beta_1 x_n+\beta_2 x_n^2+\beta_3 x_n^3+\epsilon_n,\ n=1,\dots,N, \quad \text{(Model 2)}$$
$$\epsilon_n \sim N(0,\sigma^2), n=1,\dots,N.$$

with corresponding R code
cubic_model = lm(y5 ~ x5 +I((x5)^2) + I((x5)^3), data = data_file).

If we do decide to do so, the corresponding new residuals vs. fitted values plot is



which is produced by the following code:

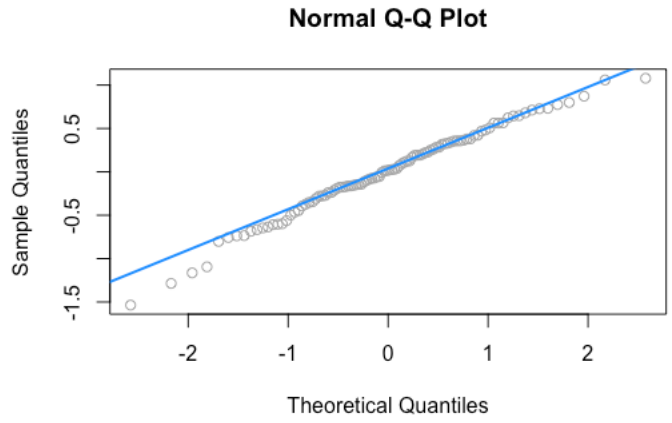
```
plot(fitted(cubic_model), resid(cubic_model), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals", main = "Data from Model 1")
abline(h = 0, col = "darkorange", lwd = 2)
```

As can be seen, this plot does appear a lot better than the previous residuals vs. fitted values plot for the basic SLR model. Next, we can also try the Breusch-Pagan test for homoscedasticity for the same cubic polynomial regression model (Model 2). From this we obtain a p-value of 0.2954, large enough to imply homoscedasticity. Furthermore, we can also try the Shapiro-Wilk test for normality of the residuals. With respect to this we obtain a p-value of 0.3937, implying normality based on the test, which is what is desired.

The Q-Q plot of the data against Model 2, produced by

```
qqnorm(resid(cubic_model), main = "Normal Q-Q Plot", col = "darkgrey")
qqline(resid(cubic_model), col = "dodgerblue", lwd = 2),
```

shows this:



The Q-Q plot displays the data aligning mostly pretty tightly along the straight (blue) line, which, based on this diagnostic tool (the Q-Q plot, that is), is good evidence that the data do follow a normal distribution, as is consistent with our regression model(s).

The output summary report corresponding to Model 2 is the following:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.53561 -0.27801  0.01787  0.35569  1.07893

Coefficients:
             Estimate Std. Error    t value Pr(>|t|)
(Intercept)  3.628e+00   2.856e-01     12.701   <2e-16 ***
x5          -4.989e-02   2.163e-02     -2.307   0.0232 *
I((x5)^2)    1.010e-03   4.509e-04      2.240   0.0274 *
I((x5)^3)    2.000e+00  2.725e-06 733863.785   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 0.5147 on 96 degrees of freedom
Multiple R-squared: 1, Adjusted R-squared: 1
F-statistic: 5.311e+13 on 3 and 96 DF, p-value: < 2.2e-16

From this, one thing that is certainly notable are the high values (values of 1) for the Multiple/Adjusted R-squared. This is of course consistent with good model accuracy. The line(s) at the top of the report concerning the residuals also do not look pretty decent, exhibiting pretty good symmetry about the median, which is close to 0 as it should be.

The F-test p-value (results at the bottom of the report) shows high significance of the overall regression model. The preponderance of the evidence we have presented here in this answer suggests that Model 2 seems quite consistent with the data. However, we can say even more.

We look also at the middle of the report, where, first, we see that the coefficient estimates for the intercept and the x^3-variable (here, I((x5)^3)) are pretty close to their true values (3 and 2, respectively) and the Residual Standard Error value is also 0.5147 close to the true value of σ in this model, which is 0.5.

We in addition see that the coefficient values of the x- and x^2-variables are both very small in absolute value (4.989e-02 and 1.010e-03) and also exhibit relatively high p-values (though a little smaller than 0.05) as compared with the other predictor variables in the model. So, we could possibly decide to simply drop them from the model altogether, which would leave us precisely with the model (Model 1) used to generate the data. Indeed, we could go through all the tests and diagnostic tools we have invoked and described here for Model 2 again but this time directly for Model 1 itself as a validation.

Problem 2: We have been finding a set of optimal parameters for the linear least-squares regression minimization problem by identifying critical points, i.e. points at which the gradient of a function is the zero vector, of the following function:

$$F(\alpha_0, \dots, \alpha_M) = \sum_{n=1}^{N} (\alpha_0 + \alpha_1 x_{n1} + \cdots + \alpha_M x_{nM} - y_n)^2 .$$

Help to justify this methodology in the following way. Letting $G: \mathbb{R}^d \to \mathbb{R}$ be any function that is differentiable everywhere, show that, if G has a local minimum at a point x_0, then its gradient is the zero vector there, i.e., $\nabla G(x_0) = \mathbf{0}$.

Solution: We assume with no real loss of generality that d=2. Consider the function G=G(u,v) in the statement of the question, which we assume has a local minimum at (u_0,v_0). We consider the partial derivative with respect to the first coordinate and, using the definition of a partial derivative, write

$$\lim_{u \to u_0} \frac{G(u,v_0) - G(u_0,v_0)}{u - u_0} = \partial_u G(u_0, v_0). \qquad \text{(B)}$$

So, if we let u approach u_0 from the left -- from below -- it follows that the denominator on the left in (B) will be negative, but the numerator must be nonnegative because (u_0,v_0) is a local minimum. So, the quotient on the left in (B) will be negative or 0. This implies that $\partial_u G(u_0, v_0) \leq 0$. But, if

we let u approach u_0 instead from the right, then a similar argument establishes that $\partial_u G(u_0, v_0) \geq 0$. Hence, $\partial_u G(u_0, v_0) = 0$. A similar argument shows that the partial derivative with respect to the other variable is 0 too. Hence the gradient of G must be 0 at the point (u_0, v_0), as was to be proved.

Problem 3: A function $G: \mathbb{R}^d \to \mathbb{R}$ is called *convex* if it satisfies the following:

$$G(\lambda u + (1-\lambda)v) \leq \lambda G(u) + (1-\lambda)G(v),$$

for any $u, v \in \mathbb{R}^d$ and λ any scalar with $\lambda \in [0,1]$. Show that the function F in Question 2 above is convex using the following steps:

a) Show that the sum of convex functions is convex.

b) Define the function

$$H(\alpha_0, \dots, \alpha_M) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_M x_M - y,$$

for any scalar, fixed constants $x_1, \dots, x_M, y$. Show that

$$H(\lambda u + (1-\lambda)v) = \lambda H(u) + (1-\lambda)H(v),$$

where u, v, and λ are as at the beginning of this question (Question 3) with $d = M + 1$.

c) Show that the function $f(t) = t^2$, t a scalar real number, is convex.

d) Explain how steps (a)-(c) can be put together to establish that the function F in Question 2 is convex.

Solution:

a)Given two convex functions G_1, G_2, consider that

$$(G_1 + G_2)(\lambda u + (1-\lambda)v) = G_1(\lambda u + (1-\lambda)v) + G_2(\lambda u + (1-\lambda)v)$$

$$\leq \lambda G_1(u) + (1-\lambda)G_1(v) + \lambda G_2(u) + (1-\lambda)G_2(v)$$

$$\leq \lambda(G_1 + G_2)(u) + (1-\lambda)(G_1 + G_2)(v),$$

b) $H(\lambda u + (1-\lambda)v) = H(\lambda u_0 + (1-\lambda)v_0, \dots, \lambda u_M + (1-\lambda)v_M)$

$$= \lambda u_0 + (1-\lambda)v_0 + \cdots + \alpha_M(\lambda u_M + (1-\lambda)v_M) - \big(\lambda + (1-\lambda)\big)y$$

$$= \lambda u_0 + \cdots + \alpha_M(\lambda u_M) - \lambda y + (1-\lambda)v_0 + \cdots + (1-\lambda)v_M) - (1-\lambda)y$$

$$= \lambda H(u) + (1-\lambda)H(v).$$

c) Choose real numbers x_1, x_2 so that x_1 and x_2 are distinct, and also select $\lambda \in (0, 1)$. Then,

$$f((1-\lambda)x_1 + \lambda x_2) = ((1-\lambda)x_1 + \lambda x_2)^2 = (1-\lambda)^2 x_1^2 + \lambda^2 x_2^2 + 2(1-\lambda)\lambda x_1 x_2 .$$

Since x_1 does not equal x_2, $(x_1 - x_2)^2 > 0$. Expanding, this means that $x_1^2 + x_2^2 > 2x_1x_2$. Hence,

$$\begin{aligned}(1-\lambda)^2x_1^2+\lambda^2x_2^2+2(1-\lambda)\lambda x_1x_2 &< (1-\lambda)^2x_1^2+\lambda^2x_2^2+(1-\lambda)(\lambda)(x_1^2+x_2^2)\\ &= (1-2\lambda-\lambda^2+\lambda+\lambda^2)\,x_1^2+(\lambda-\lambda^2+\lambda^2)\,x_2^2\\ &= (1-\lambda)\,x_1^2+\lambda x_2^2\\ &= (1-\lambda)f(x_1)+\lambda f(x_2)\,.\end{aligned}$$

d) With H the function from Problem 3b, let H^2 be the function defined by

$$H^2(u) = (H(u))^2 \text{ for } u \in \mathbb{R}^{M+1}.$$

Note that, by repeated use of Part 3a, convexity of the function F will follow if we can show that H^2 is a convex function for any choice of the fixed values $x_1, \dots, x_M, y$. But, consider that

$$\begin{aligned}H^2(\lambda u + (1-\lambda)v) &= \big(\lambda H(u) + (1-\lambda)H(v)\big)^2 \text{ (by Part 3b)}\\ &\le \lambda H^2(u) + (1-\lambda)H^2(v) \text{ (by Part 3c),}\end{aligned}$$

which implies the convexity of H^2.

Problem 4: Show that if a convex function $G: \mathbb{R}^d \to \mathbb{R}$ has a local minimum at a point x_0, then this local minimum is also a global minimum for the function over all of its domain $\mathbb{R}^d$.

Solution: Since x_0 is a local minimum, for any $y \in \mathbb{R}^d$, we can choose a small enough $\alpha > 0$, such that

$$G(x_0) \le G(x_0 + \alpha(y - x_0)). \qquad (C)$$

Furthermore, since G is a convex function, we have

$$G(x_0+\alpha(y-x_0))=G(\alpha y+(1-\alpha)x_0)\le\alpha G(y)+(1-\alpha)G(x_0). \qquad (D)$$

Combining (C) and (D), we have

$$G(x_0) \le \alpha G(y) + (1-\alpha)G(x_0),$$

which implies that $G(x_0) \le G(y)$. Since y is an arbitrary point in $\mathbb{R}^d$, this proves

that x_0 is a global minimum.

Problem 5: Obtaining data samples from the dataset "HW 3 Dataset 2" (which will be sent in an email with that subject line), utilize any or all of the various regression model-fit assessment visual plots and/or graphs, results of statistical hypothesis tests, and numerical regression model accuracy measures, as well as any other applicable diagnostic tools available in R that we have discussed in the course thus far, to help identify and validate a linear regression model that appears to best fit the given data. These will likely include for example direct X-Y variable graphs and residuals vs. fitted value plots implemented in R. You will likely need to try a number of different response and/or predictor variable transformations, for example, using the diagnostic tools to decide which regression models appear consistent with the data and which do not. You should identify one or perhaps two candidate regression models that you believe are most consistent with the given data. Please try to include screenshots of the graphical plots you use as well as quote any relevant R code output. You can or even should include the R code itself as well if you feel it helps to support your argument. Please explain your reasoning as to why the model(s) you propose may be the right one(s) in plain natural language prose.

Solution:
The R code used to generate the data for this problem is the following:

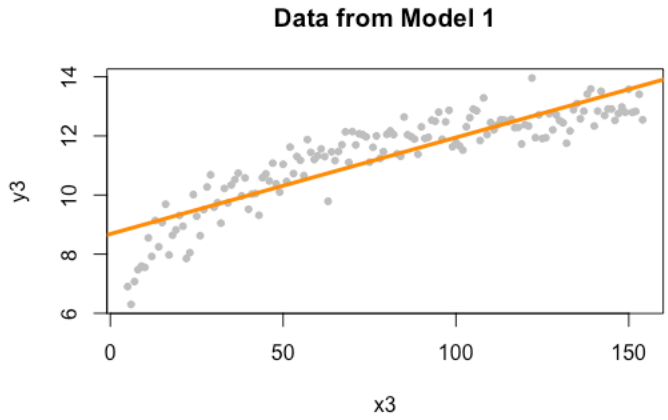
```
x3 <- seq(5,154)
e3 <- rnorm(150, mean = 0,  sd = 0.5)
y3 <- 3+2*log(x3)+e3

data_file <- data.frame(y3,x3)
```

This corresponds to the (instantiated) regression model

$$Y_n = 3 + 2\log(x_n) + \epsilon_n,\ n = 1, \ldots, N, \qquad \text{(Model 1)}$$
$$\epsilon_n \sim N(0, 0.25), n = 1, \ldots, N,$$

with N=150.

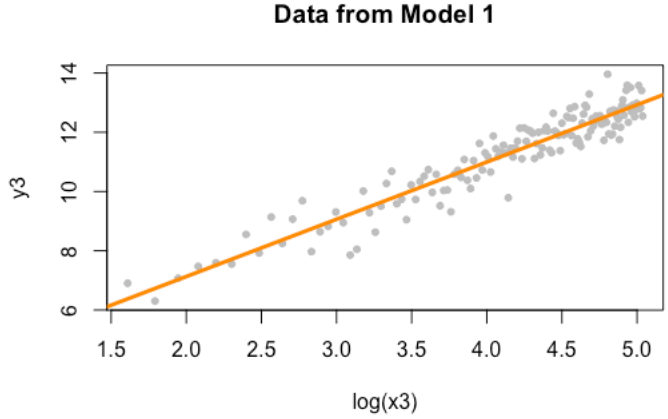
Note that the direct Predictor vs. Response (X-Y variable) plot of the generated data along a regression line according to a Simple Linear Regression model looks like the following graph:



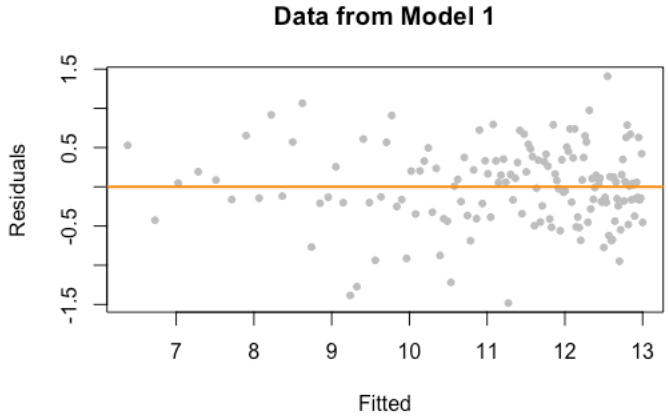
The second graph above is of course the Residuals vs. Fitted Values plot. Both plots above strongly suggest that the standard SLR model (without any variable transformation) is not consistent with this dataset. Yet if we try the class of candidate models

$$Y_n=\beta_0 + \beta_1 \log(x_n) + \epsilon_n,\ n = 1, \ldots, N, \qquad \text{(Model 2)}$$
$$\epsilon_n \sim N(0, \sigma^2), n = 1, \ldots, N,$$

the corresponding Y- log (X) plot looks pretty good:



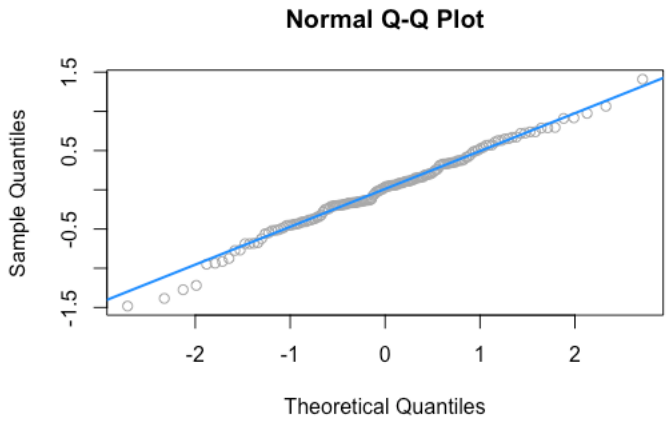
Moreover the corresponding residuals vs fitted values plot for its part shows dramatic improvement relative to, for example, the SLR model:



Next, we try the Breusch-Pagan test for homoscedasticity for the same cubic polynomial regression model (Model 2). From this we obtain a p-value of 0.1444, large enough to imply homoscedasticity. Furthermore, we can also try the Shapiro-Wilk test for normality of the residuals. With respect to this we obtain a p-value of 0.4953, implying normality based on the test, which is what is desired.

Note, for example, that the fitted vs. residuals plot on the previous page does not show an even distribution across the (orange) 0 line -- the data points are almost all above it in the center but below it on the sides. In this sense the fitted vs. residuals plot above on this page looks significantly better. Moreover, the results of the Breusch-Pagan test, in particular, help to support this view.

The Q-Q plot corresponding to Model 2 with the data in question also looks quite favorable, showing this:



Finally, we have the summary output report for candidate Model 2:

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.48239 -0.31379  0.03454  0.33745  1.40947

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2640     0.2207   14.79   <2e-16 ***
log(x3)       1.9329     0.0523   36.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5017 on 148 degrees of freedom
Multiple R-squared:  0.9022,  Adjusted R-squared:  0.9016
F-statistic:  1366 on 1 and 148 DF,  p-value: < 2.2e-16
```

One thing we see from this report, once again, is the favorable (relatively high) values for the Multiple R-squared and Adjusted R-squared.

We see that these diagnostic tools which we have discussed in the course are validated in that they give results quite consistent with what we would expect when applied against a given regression model along with data generated by means of that known model.