PSTAT 126 Homework #2

P1) Explain why the coefficient of determination R^2 satisfies $0 \leq R^2 \leq 1$. Is the same true of the adjusted coefficient of determination R_a^2? Give a reason.

The coefficient of determination, R^2, is a goodness of fit measure, where it examines how the differences in 1 variable can be explained by the difference in the 2nd variable. It focuses on the linear relationship between 2 variables. It is between 0 and 1, a value of 1 is a perfect fit and suggests it is a reliable model, while a value of 0, suggests the model fails to accurately model the data.

Yes, the same is true for R_a^2, they are both used as measures of regression model accuracy because $0 \leq R_a^2 \leq 1$, where the model was adjusted with a factor: $\frac{N-1}{N-M-1}$ to correct the trend of increasing with more predictor variables in R^2. The model was adjusted but the values of 0 and 1 still represent the same concepts.

P2) Show, in the context of Simple Linear Regression, that the residuals e_n, $n = 1,\ldots,N$, are normally-distributed if the noise/error terms $\epsilon_n, n=1,\ldots,N$, are. Give your reasoning.

SLR: $Y_n = \beta_0 + \beta_1 x_n + \varepsilon_n$, $n=1,\ldots,N$ where $\varepsilon_n \sim N(0,\sigma^2)$, $n=1,\ldots,N$

where Y_n is Normally distributed and error terms are as well while the residuals are.

$$e_n := y_n - \hat{y}_n \text{ where } \hat{y}_n := \hat{\beta}_0 + \hat{\beta}_1 x_n, n=1,\ldots,N$$

$$e_n = (\beta_0 + \beta_1 x_n + \varepsilon_n) - (\hat{\beta}_0 + \hat{\beta}_1 x_n)$$

$$e_n = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x_n + \varepsilon_n$$

e_n is a linear function of ε_n and a linear combination of independent normally-distributed R.V.'s are also normally distributed. So, e_n is normally distributed.

P3) For any given, fixed number N of samples, explain why and in what sense each $\hat{\beta}_m$, $m = 0, \ldots, M$, is an optimal estimator for β_m, $m = 0, \ldots, M$, respectively.

By the Gauss-Markov Thm,

1) $E[\hat{\beta}_m] = \beta_m$ meaning the estimators $\hat{\beta}_m$ for the regression paramets β_m are unbiased

2) $\hat{\beta}_m$ are of min. var. amongst unbiased linear estimates for β_m.

This implies, α_m (unbiased linear estimators), the error $E[(\alpha_m - \beta_m)^2] = E[\alpha_m - E[\alpha_m]^2]$ is minimized when $\alpha_m = \hat{\beta}_m$.

So, $\hat{\beta}_m$ are optimal estimators for a fixed # N of samples

P4) In R, use the lm( ) function to generate a summary output report with Temp as the response against the predictors Ozone, Wind, and Month, using the "built-in" airquality dataset. Is the regression model overall a significant one that adds insight beyond a simple "intercept-only" model? How do you know? Which of the three predictor variables are deemed important for the regression model? How do you know this? Based on the summary report, what do the residuals say about the validity of the regression model?

Yes, the regression model is significant because the p-value $< 2.2e^{-16}$ is high. It also implies the significance of the SLR model w/both parameters: intercept & slope.

Ozone is deemed important for the regression model because it's t-value is large, and a larger value implies more confidence in the corresponding parameter estimate. while month would also be after Ozone and Wind would not be since it's t-value is very low.

The residuals are Normally distributed since 1Q and 3Q are balanced and the median is close to 0. The RSE is 6.159 on 112 degrees of freedom, where the smaller RSE indicates a better model fit, thus the model is valid.

P5) With the built-in mtcars dataset, taking mpg as the Y-variable and disp and hp as the x-variables, use R to solve the resulting least squares multiple regression problem for the vector $\hat{\mathbb{B}}$ (see course slides 31-32) by direct computation of the matrix product

$$(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y},$$

assuming invertibility of the matrix $\mathbb{X}^T\mathbb{X}$. To do this, use the R code:

```
n = nrow(mtcars)
p = length(coef(mtcars))
X = cbind(rep(1, n), mtcars$disp, mtcars$hp)
y = mtcars$mpg

(beta_hat = solve(t(X) %*% X) %*% t(X) %*% y)
```

Do you know, using R, another way to do this computation and generate the vector $\hat{\mathbb{B}}$? Do do the computation to produce this vector a different way. Include your code in your answer, and compare the results with those of the other method.

Using given code:

```
> n= nrow(mtcars)
> p = length(coef(mtcars))
> X = cbind(rep(1, n), mtcars$disp, mtcars$hp)
> y = mtcars$mpg
> (beta_hat = solve(t(X) %*% X) %*% t(X) %*% y)
           [,1]
[1,] 30.73590425
[2,] -0.03034628
[3,] -0.02484008
```

$\hat{\mathbb{B}} = (30.73590425, -0.03034628, -0.02484008)$

Another way to compute:

```
> data = lm(mpg~disp+hp,data=mtcars)
> coef(data)
(Intercept)        disp          hp
30.73590425 -0.03034628 -0.02484008
>
```

$\hat{\mathbb{B}} = (30.73590425, -0.03034628, -0.02484008)$

The parameter estimates are the same with utilizing the matrix method or lm() function in R

P6) Do Problem 2.13 in the Weisberg (2014) text. Note that "regression of dheight on mheight" in the first part means that mheight is the predictor variable. response

dheight ~ predictor mheight

2.13 Heights of mothers and daughters (Data file: `Heights`)

2.13.1 Compute the regression of `dheight` on `mheight`, and report the estimates, their standard errors, the value of the coefficient of determination, and the estimate of variance. Write a sentence or two that summarizes the results of these computations.

2.13.2 Obtain a 99% confidence interval for β_1 from the data.

2.13.3 Obtain a prediction and 99% prediction interval for a daughter whose mother is 64 inches tall.

2.13.1

The estimated values are $\hat{\beta}_0 = 29.91744$, $\hat{\beta}_1 = 0.54175$, the Std. Errors are $\hat{\beta}_0 = 1.62247$, $\hat{\beta}_1 = 0.02596$

The coeff. of determination, $R^2 = 0.2408$ and the RSE is 2.266. This is used to calculate the estimate of variance, you square it $(2.266)^2$ to calculate the estimate of variance: 5.134756 The regression model is significant since the p-value $< 2.2e^{-16}$ is high.

2.13.2 99% confint

	0.5%	99.5%
$\hat{\beta}_1$ =	0.4747836	0.6087104

2.13.3

	fit	lwd	upr
	64.58925	58.74045	70.43805

P7) Do Problem 2.15 in the Weisberg (2014) text.

2.15 Smallmouth bass (Data file: `wblake`)

2.15.1 Using the West Bearskin Lake smallmouth bass data in the file `wblake`, obtain 95% intervals for the mean length at ages 2, 4, and 6 years.

2.15.2 Obtain a 95% interval for the mean length at age 9. Explain why this interval is likely to be untrustworthy.

2.15.1

Mean Length @ age 2:

	fit	lwr	upr
	126.17494	69.73151	182.61837

Mean Length @ age 4:

	fit	lwr	upr
	186.8227	130.4572	243.1882

Mean Length @ age 6:

	fit	lwr	upr
	247.4705	191.0533	303.8877

2.15.2

Mean Length @ age 9:

	fit	lwr	upr
	338.4422	281.7056	395.1788

The interval is likely to be untrustworthy b/c the prediction (fit) is too high for the mean length to be between 281.7056 to 395.1788.

P8) Do Problem 2.17.1 in the Weisberg (2014) text. With respect to Problem 2.17.2, simply compute $\hat{\beta}_1$, but you can ignore the rest of it (Hint: Models are fit in R without the intercept by adding a -1 to the formula. Also, you do not need to do Problem 2.17.3.)

2.17 Regression through the origin Occasionally, a mean function in which the intercept is known a priori to be 0 may be fit. This mean function is given by

$$\mathrm{E}(y|x) = \beta_1 x \tag{2.27}$$

The residual sum of squares for this model, assuming the errors are independent with common variance σ^2, is $\mathrm{RSS} = \sum\left(y_i - \hat{\beta}_1 x_i\right)^2$. ① ②

2.17.1 Show that the least squares estimate of β_1 is $\hat{\beta}_1 = \sum x_i y_i / \sum x_i^2$. Show that $\hat{\beta}_1$ is unbiased and that $\mathrm{Var}\left(\hat{\beta}_1|X\right) = \sigma^2 / \sum x_i^2$. Find an expression for $\hat{\sigma}^2$. How many df does it have? ③

$\hat{\beta}_1 = \dfrac{\sum x_i y_i}{\sum x_i^2}$ unbiased

$\mathrm{var}\left(\dfrac{\hat{\beta}_1}{X}\right) = \dfrac{\sigma^2}{\sum x_i^2}$

① $\dfrac{\partial RSS}{\partial \beta_1} = \left[\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i)\right] x_i = 0$

$\sum_{i=1}^{n}(y_i x_i - \hat{\beta}_1 x_i^2) = 0$

$\sum y_i x_i = \hat{\beta}_1 \sum x_i^2$

$\hat{\beta}_1 = \dfrac{\sum x_i y_i}{\sum x_i^2}$

② $E(\hat{\beta}_1) = \dfrac{1}{\sum x_i^2} \sum x_i E(y_i)$

$= \dfrac{1}{\sum x_i^2} \sum x_i (x_i \beta_1)$

$= \beta_1$

where $\hat{\beta}_1$ is an unbiased estimator of β_1

③ $\mathrm{var}(\hat{\beta}_1) = \dfrac{\sum_{i=1}^{n} x_i^2 \mathrm{var}(y_i)}{\left(\sum x_i^2\right)^2}$

$= \dfrac{\sigma^2 \sum x_i^2}{(\sum x_i^2)^2}$

$\mathrm{var}(\hat{\beta}_1|X) = \dfrac{\sigma^2}{\sum x_i^2}$

④ $\hat{\sigma}^2 = \dfrac{\sum (y_i - \hat{y}_1)^2}{df}$

$= \dfrac{\sum (y_i - \hat{\beta}_1 x_i)^2}{n-1}$ → $(y_i - \hat{\beta}_1 x_i)(y_i - \hat{\beta}_1 x_i)$

$= y_i^2 - 2 y_i \hat{\beta}_1 x_i + \hat{\beta}_1^2 x_i^2$

$= \dfrac{\sum y_i^2 - 2\hat{\beta}_1 \sum x_i - \sum x_i^2}{n-1}$

$df = n-1$

since there is only 1 parameter β_1