PSTAT 126: Homework #2 Solutions

Problem 1: 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.

Solution: $R^2 = 1 - \frac{\sum_{n=1}^N e_n^2}{\sum_{n=1}^N (y_n - \bar{y})^2} = 1 - \frac{\sum_{n=1}^N (y_n - \hat{y}_n)^2}{\sum_{n=1}^N (y_n - \bar{y})^2}$, where $\bar{y} = \frac{1}{N}\sum_{n=1}^N y_n$.

So, by what we are calling the Decomposition of Variation formula, i.e.,

$\sum_{n=1}^N (y_n - \bar{y})^2 = \sum_{n=1}^N (y_n - \hat{y}_n)^2 + \sum_{n=1}^N (\bar{y} - \hat{y}_n)^2$,

we find that

$$R^2 = 1 - \frac{\sum_{n=1}^N (y_n - \bar{y})^2 - \sum_{n=1}^N (\bar{y} - \hat{y}_n)^2}{\sum_{n=1}^N (y_n - \bar{y})^2}$$

$$= \frac{\sum_{n=1}^N (\bar{y} - \hat{y}_n)^2}{\sum_{n=1}^N (y_n - \bar{y})^2} \geq 0.$$

Then, another appeal to the Decomposition of Variation formula ensures that $R^2 \leq 1$ as well.

Moreover, no, the analogous assertion is not true in general for R_a^2, which satisfies $R_a^2 = 1 - (1 - R^2)\frac{N-1}{N-M-1}$, where *N* is the number of samples and *M* is the number of x-variables. In fact, we can see from this definition that for small R^2 and large enough *N* with *M* less than but close to *N* (for example, consider *M=N-2, N>2,* $R^2 < 0.5$), R_a^2 should be negative. We note, however, that, technically speaking, for the complete counterexample we should ideally exhibit a full, viable linear regression model in which all of these values for R^2,*N*, and *M* rendering R_a^2 negative are realized simultaneously.

Problem 2: 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 ϵ_n,n=$1,\ldots,N$,are. Give your reasoning.

Solution: We have $e_n = y_n - \hat{y}_n$. We know the y_n are normally-distributed because we are assuming the noise/error terms are. We also know the general fact that linear combinations of independent, normally-distributed random variables are themselves normally-distributed.

$\hat{\beta}_1 = \frac{\sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=1}^N (x_n - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$, $\hat{\beta}_0 = \frac{1}{N}\left(\sum_{n=1}^N y_n - \hat{\beta}_1 \sum_{n=1}^N x_n\right)$,

where $\bar{y} = \frac{1}{N}\sum_{j=1}^N y_j$ and similarly for $\bar{x}$, and the x_n can be regarded as non-random, fixed values. We insert these two equations into $e_n = y_n - \hat{y}_n$ using the definition of $\hat{y}_n$ seen in the lecture slides and then collect coefficients of like y_n terms. Then it is not hard to see that the fact we cited about linear combinations of independent normally distributed random variables will complete the answer.

Problem 3: For any given, fixed number N of samples, explain why and in what sense each $\hat{\beta}_m$, m = 0, ...,M, is an optimal estimator for β_m, m = 0,...,M, respectively.

Solution: You can use the statement of the Gauss-Markov Theorem for this. First that theorem tells us that the $\hat{\beta}_m$, $m = 0, \dots, M$, are unbiased estimators for the β_m, $m = 0, \dots, M$, respectively. So $\boldsymbol{E}[\hat{\beta}_m]= \beta_m$. The theorem also states that the $\hat{\beta}_m$ are of minimum variance among all unbiased, linear estimators for β_m. But, this then implies as well that, among all unbiased, linear estimators $\hat{\alpha}_m$, the error $\boldsymbol{E}[(\hat{\alpha}_m - \beta_m)^2]$ = $\boldsymbol{E}[(\alpha_m - \boldsymbol{E}[\hat{\alpha}_m])^2]$ is minimized when $\hat{\alpha}_m = \hat{\beta}_m$. This says that taking $\hat{\alpha}_m = \hat{\beta}_m$ minimizes the expected square of the difference between $\hat{\alpha}_m$ and β_m among all unbiased linear estimators $\hat{\alpha}_m$ for β_m, and hence in this important sense is optimal.

Problem 4: 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?

Solution: The associated Summary output report is the following:

```
Call:

lm(formula = Temp ~ Ozone + Wind + Month, data = airquality)

Residuals:

    Min      1Q  Median      3Q     Max

-21.4801  -4.2884   0.4907   4.6106  12.4071

Coefficients:

            Estimate  Std. Error  t value  Pr(>|t|)

(Intercept) 58.97480    4.10572   14.364   < 2e-16 ***

Ozone        0.17060    0.02183    7.816   3.22e-12 ***

Wind        -0.24236    0.20302   -1.194    0.235
```

Month   1.95866   0.39830   4.918   3.03e-06 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.159 on 112 degrees of freedom

  (37 observations deleted due to missingness)

Multiple R-squared: 0.5893, Adjusted R-squared: 0.5783

F-statistic: 53.58 on 3 and 112 DF, p-value: < 2.2e-16

Now, the F-test p-value in the last line of this report is very small (2.22e-16). This means we reject the null hypothesis and the overall regression is significant. The t-test results in the middle of the report show which individual x-variables are significant in the regression. Specifically, we look at the values corresponding to the three predictor variables under “Pr(>|t|)”. So the p-values for Ozone and Month are small – well below 0.05 – so we interpret this as implying that these variables are significant. However, we see that Wind is not. The residuals report suggests that the distribution of the residuals appears to be fairly well-balanced at least in so far as the limited information in the report goes. For example, ideally, the 1Q value should be the negative of the 3Q value (this appears to be close to being true in the case of this report) and the median should be close to 0 as well (this is also reasonably close to being true in this case).

Problem 5: 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 $\widehat{\mathbb{B}}$ (see course slides 37-38) by direct computation of the matrix product

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

assuming invertibility of the matrix $\mathbb{X}^{\mathrm{T}}\mathbb{X}$. 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 $\widehat{\mathbb{B}}$? 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.

Solution:

```
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)

mpg_model = lm(mpg ~ wt + year, data = autompg)
coef(mpg_model)

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

            [,1]
[1,] 30.73590425
[2,] -0.03034628
[3,] -0.02484008

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

Problem 6: 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.

Solution:

2.13.1) The information is contained in the corresponding summary output report:

Call:

lm(formula = dheight ~ mheight, data = econ13)

Residuals:

Min 1Q Median 3Q Max

-7.397 -1.529 0.036 1.492 9.053

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 29.91744   1.62247   18.44   <2e-16 ***

mheight   0.54175   0.02596   20.87   <2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.266 on 1373 degrees of freedom

Multiple R-squared: 0.2408, Adjusted R-squared: 0.2402

F-statistic: 435.5 on 1 and 1373 DF, p-value: < 2.2e-16

The t-statistic for the slope has a p-value very close to 0, suggesting strongly that β_1 does not equal 0. The value of R^2 = 0.2408, so only about one-fourth of the variability in daughter's height is explained by mother's height.

2.13.2) confint(Heights, level = 0.99)

                    0.5 %   99.5 %

mheight   0.4747836 0.6087104

2.13.3) The R code

height_model2 = lm(dheight ~ mheight, data = Heights)

predict(height_model2, data.frame(mheight=64), interval="prediction", level=.99)

produces the answer

fit   lwr   upr

64.58925   58.74045 70.43805

Problem 7: Do Problem 2.15 in the Weisberg (2014) text.
Solution:
2.15.1) The R code

```
model1 = lm(Length ~ Age, wblake)
new_speeds = data.frame(Age = c(2,4,6))
predict(model1, newdata = new_speeds,
       interval = c("confidence"), level = 0.95)
```

produces

```
   fit      lwr      upr
1 126.1749 122.1643 130.1856
2 186.8227 184.1217 189.5237
3 247.4705 243.8481 251.0929
```

2.15.2) For Age 9, we have

```
fit      lwr      upr
1 338.4422 331.4231 345.4612
```

This is an extrapolation outside the range of the data, as there were no fish older than 8 years in the sample. We do not really know if the straight-line mean function truly applies at age 9.

Problem 8: 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.)

Solution:

2.17.1) Note that the model in this problem is just like a Simple Linear Regression model, but with $\beta_0 = 0$. Take the derivative with respect to $\alpha_1 = \hat{\beta}_1$ in a way analogous to the solution of HW #1, Problem 6:

$$0 = \frac{d}{d\alpha_1}\left(\sum_{n=1}^{N} (y_n - \alpha_1 x_n)^2\right)\Big|_{\alpha_1=\hat{\beta}_1}$$

$$= 2\sum_{n=1}^{N} x_n\left(y_n - \hat{\beta}_1 x_n\right).$$

Hence,

$$\sum_{n=1}^{N} y_n\, x_n = \sum_{n=1}^{N} \hat{\beta}_1 x_n^2,$$

and we can then solve for $\hat{\beta}_1$. Next,

$E(\hat{\beta}_1) = E\left(\sum_{n=1}^{N} Y_n\, x_n / \sum_{n=1}^{N} x_n^2\right)$

$= \sum_{n=1}^{N} x_n E(Y_n) / \sum_{n=1}^{N} x_n^2$ (this follows because the x_n can be presumed to be fixed constants)

$= \sum_{n=1}^{N} x_n(\beta_1 x_n) / \sum_{n=1}^{N} x_n^2 = \beta_1$.

$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\sum_{n=1}^{N} Y_n\, x_n / \sum_{n=1}^{N} x_n^2\right)$

$= \sum_{n=1}^{N} x_n^2 \mathrm{Var}(Y_n) / \left(\sum_{n=1}^{N} x_n^2\right)^2$

$= \sigma^2 \sum_{n=1}^{N} x_n^2 / \left(\sum_{n=1}^{N} x_n^2\right)^2$

$= \sigma^2 / \sum_{n=1}^{N} x_n^2$

The variance σ^2 is the variance of each random variable Y_n, that is, $\mathrm{Var}(Y_n)$, for any $n = 1, \dots, N$. $\mathrm{Var}(Y_n)$, is of course defined as $E[(Y_n - E[Y_n])^2]$, where in the case of general simple linear regression we have $E[Y_n]$ =$\beta_0 + \beta_1 x_n$, but in this problem we assume $\beta_0 = 0$, so that here $E[Y_n]$ =$\beta_1 x_n$. So, in order to get an estimate of σ^2 we want an estimator for $(Y_n - E[Y_n])^2$, hence, in turn, what we want is an estimator of $E[Y_n]$ =$\beta_1 x_n$. In the case of general simple linear regression our estimate of $E[Y_n]$ =$\beta_0 + \beta_1 x_n$ is the fitted value $\hat{\beta}_0 + \hat{\beta}_1 x_n$. Here in this problem, again, we assume $\beta_0 = 0$, so our approximation for $E[Y_n]$ =$\beta_1 x_n$ is $\hat{\beta}_1 x_n$. Thus our estimator $\hat{\sigma}^2$ for σ^2 in the setting of this problem is the average $\sum_{n=1}^{N}\left(Y_n - \hat{\beta}_1 x_n\right)^2$ over N summands, but we want to multiply this sum by a suitable constant factor so that $\hat{\sigma}^2$ will be an unbiased estimator for σ^2. According to Sec. 2.3 of the Weisberg (2014) text, we obtain the unbiased estimator $\hat{\sigma}^2$ by dividing $\sum_{n=1}^{N}\left(Y_n - \hat{\beta}_1 x_n\right)^2$ by the number of degrees of freedom in this setting, which in this context is the number of cases N minus the number of parameters in the mean function, in this case just one. Hence our estimator for σ^2 is

$$\hat{\sigma}^2 = \left(\frac{1}{N-1}\right)\sum_{n=1}^{N}\left(Y_n - \hat{\beta}_1 x_n\right)^2.$$

2.17.2) The R code

```
snake_model = lm(Y ~ X-1, data = snake)
summary(snake_model)
```

gives $\hat{\beta}_1 = \ 0.52039$.