Q.No : 1)

[10 points] [L3, CO 1] A simple linear regression model for predicting length of gestation in mammals (in number of days until birth) using birth weight (in Kg) is built as follows:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 187.0837    26.9426   6.944 6.73e-05 ***
Birthwgt      3.5914     0.5247   6.844 7.52e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.09 on 9 degrees of freedom
Multiple R-squared:  0.8388,    Adjusted R-squared:  0.8209
F-statistic: 46.84 on 1 and 9 DF,  p-value: 7.523e-05
```

(a) Estimate the gestation period of a mammal that would give birth to an offspring that weighs 3.5Kg.

(b) For a 1Kg increase in birth weight, we can expect the gestation period to increase/decrease (choose one) by what amount?

(c) Interpret the predicted gestation period for a 0Kg birth weight. Is it meaningful?

(d) How accurate is the model?

(e) For a 1Kg decrease in birth weight, we can be 95% confident that the length of gestation will increase/decrease (choose one) by at least ____ days and at most ____ days.

Page:1

$\hat{y} = 187.0837 + 3.5914 * \text{Birthwgt}$

(a) birthwgt = 3.5 kg

$\hat{y} = 187.0837 + 3.5914 * 3.5$

$\hat{y} = 199.6536$ days

(b) For a 1 kg increase in birth weight, we can gestation period to increase by 3.5914 days

OR, With 95 % confidence, we can say that -ion period will increase by 3.5 to 4.5 days increase in the birthwgt.

(c) $\hat{y} = \hat{\beta}_0 = 187.0837$ days

For a 0 kg birth weight, the gestation period is ately 187 days ; which is not meaningful.

But, we can also interpret as 'The minimum -on period of a mammal is around 187 days'. W somewhat meaningful.

(d) The model is 82.09 % Accurate.

(e) For a 1 kg decrease in birth weight, we can confident that the length of gestation will dec at least 2.542 days and at most 4.6408 day

Q.No : 2)

[10 points] [L3, CO 1] 3. Suppose we fit a linear regression model to estimate credit card balance (in dollars) as a function of ethnicity (African-American, Asian, and Caucasian) resulting in the following:

	Coefficient	Std. error	t-statistic	p-value
Intercept	531.00	46.32	11.464	< 0.0001
ethnicityAsian	−18.69	65.02	−0.287	0.7740
ethnicityCaucasian	−12.50	56.68	−0.221	0.8260

(a) What is the predicted credit balance for each ethnicity?

(b) Briefly interpret the values of the regression coefficient estimates in the context of credit card balance and ethnicity.

(c) What does this model suggest about the difference in credit card balance between the ethnicities? Be short and precise with your answer.

Page:1

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 *$ ethnicity Asian $+ \hat{\beta}_2 *$ ethnicity Cauc

(a) Predicted credit balance for African_Americ

$\hat{y}_{A-A} = \hat{\beta}_0 = 531.00$ dollars

Predicted credit balance for Asian ;

$\hat{y}_{Asian} = \hat{\beta}_0 + \hat{\beta}_1 *$ ethnicity Asian

$= 531 + (-18.69)(1)$

$\hat{y}_{Asian} = 512.31$ dollars

Predicted credit balance for Caucasian ;

$\hat{y}_C = \hat{\beta}_0 + \hat{\beta}_2 *$ ethnicity Caucasian

$= 531 + (-12.5)(1)$

$\hat{y}_C = 518.5$ dollars

(b) The regression coefficient estimate $\hat{\beta}_0$ sugg

the credit card balance of African-Americans

-und 531 dollars.

$\hat{\beta}_1$ suggests that the credit card balance of A

is around 18.69 dollars lesser than that of A

Americans.

$\hat{\beta}_2$ suggests that the credit card balance of Ca
is 12.5 dollars lesser than that of African-Ame

Page:2

(c) From the model, observing the 'p-values' for asian and ethnicity Caucasian, which are infact far than the significant Threshold, we cannot reject - hypothesis of presence of non-linear relationship b credit card balance and ethnicity.

Q.No : 3)

[10 points] [L2, CO 1] A multiple linear regression model for predicting house price (in dollars) as a function of living area (square feet) and type of fuel used for heating (a categorical variable) is built as follows:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8411.608   5538.298   1.519  0.12899
livingArea    110.231      2.784  39.590  < 2e-16 ***
fuelgas     14630.087   4530.883   3.229  0.00127 **
fueloil      -252.581   6111.020  -0.041  0.96704
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68830 on 1724 degrees of freedom
Multiple R-squared:  0.5119,    Adjusted R-squared:  0.5111
F-statistic: 602.8 on 3 and 1724 DF,  p-value: < 2.2e-16
```

(a) What is the name of the categorical variable before dummy encoding?

(b) How many levels does the categorical variable have?

(c) Identify the reference level for the categorical variable (pick one): solar, thermal, motor, electric, generator, wind, tidal.

(d) What are the non-reference levels of the categorical variable?

(e) What is the predicted house price of a house with neither gas nor oil as fuel for heating?

Page:1

(a) name of categorical variable before dummy enco

fuel.

(b) categorical variable have 3 levels.

(c) Reference level for categorical variable is el

(d) Non-reference levels of categorical variable are

fuel gas and fuel oil.

(e) $\hat{y}_e = \hat{\beta}_0 + \hat{\beta}_1 * \text{livingArea}$

$\hat{y}_e = 8411.608 + 110.231 * \text{livingArea}$ do

Q.No : 4)

[10 points] [L3, CO 1] Suppose we are interested in a linear model for predicting instructor evaluation score as a function of age and gender. Assume there are two genders: female and male. The output of fitting a model that captures the interaction between age and gender is shown below:

term	estimate	std_error	statistic	p_value
intercept	4.883	0.205	23.80	0.000
age	-0.018	0.004	-3.92	0.000
gendermale	-0.446	0.265	-1.68	0.094
age:gendermale	0.014	0.006	2.45	0.015

Write down the predicted instructor evaluation scores for a male and female instructor; simplify as much as possible. Interpret the effect of age on instructor evaluation score for both genders.

Page:1

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * age + \hat{\beta}_2 * gendermale + \hat{\beta}_3 * age$$

For a male instructor;

$$\hat{y}_{male} = \hat{\beta}_0 + \hat{\beta}_1 * age + \hat{\beta}_2 * 1 + \hat{\beta}_3 * age *$$

$$\hat{y}_{male} = \hat{\beta}_0 + \hat{\beta}_2 + (\hat{\beta}_1 + \hat{\beta}_3) * age$$

$$\hat{y}_{male} = 4.883 + (-0.446) + (-0.018 + 0.014)$$

$$\boxed{\hat{y}_{male} = 4.437 - 0.004 * age}$$

For a Female instructor;

$$\hat{y}_{female} = \hat{\beta}_0 + \hat{\beta}_1 * age$$

$$\hat{y}_{female} = 4.883 + (-0.018) * age$$

$$\boxed{\hat{y}_{female} = 4.883 - 0.018 * age}$$

$$\hat{y} \quad - 0.004 * age$$

$\hat{y}_{male} = 4.454 - 0.008 * age$

$\hat{y}_{female} = 4.883 - 0.018 * age$

From the predicted instructor evaluation scores and female instructors, we can say the following

Page:2

→ For a male instructor, we can say that as increases by 1 unit, the evaluation score wi - ase by 0.004 units.

→ similarly, for a female instructor, as the increase by 1 unit, the evaluation score d down by 0.018 units.

Q.No : 5)

[10 points] [L5, CO 1] A multiple linear regression model for how much air (in liters) a child can forcefully exhale from the lungs, referred to as the forced exhalation volume (FEV), as a function of height (in inches) and gender (female and male) is built as follows:

```
lm(formula = FEV ~ height + gender + height:gender, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54654 -0.25282  0.00649  0.25666  2.00491

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -4.318219   0.297637 -14.508  < 2e-16 ***
height            0.112426   0.004928  22.815  < 2e-16 ***
gendermale       -1.545629   0.373843  -4.134 4.02e-05 ***
height:gendermale  0.027457   0.006119   4.487 8.54e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4204 on 650 degrees of freedom
Multiple R-squared:  0.766,     Adjusted R-squared:  0.7649
F-statistic: 709.2 on 3 and 650 DF,  p-value: < 2.2e-16
```

(a) Write down separate equations for predicting FEV as a function of height for female and male children. Clearly show the coefficient estimates in the equations.

(b) For a 1inch increase in height, does the predicted FEV increase or decrease (choose one) for female and male children?

(c) Select which gender is predicted to have a higher increase in FEV for a unit increase in height?

Page:1

(a) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * \text{height} + \hat{\beta}_2 * \text{gendermale} + \hat{\beta}_3 * \text{gend}$

For female children;

$\hat{y}_f = \hat{\beta}_0 + \hat{\beta}_1 * \text{height}$

$\hat{y}_f = -4.3182 + 0.1124 * \text{height}$

For male children;

$\hat{y}_m = \hat{\beta}_0 + \hat{\beta}_1 * \text{height} + \hat{\beta}_2 + \hat{\beta}_3 \text{ height}$

$\hat{y}_m = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) * \text{height}$

$\hat{y}_m = (-4.31822 - 1.54563) + (0.11243 + 0.0$

$\hat{y}_m = -5.86385 + 0.13983 * \text{height}$

(b) For 1 inch increase in height, the predicte increases for both males and females.

(C) $\hat{y}_{female} = -4.3182 + 0.1124 * \text{height}$

$$\hat{y}_m = -5.86385 + 0.13983 * \text{height}$$

By comparing the co-efficient estimates of height male and female children, we can say male are predicted to have a higher increase in a unit increase in height.