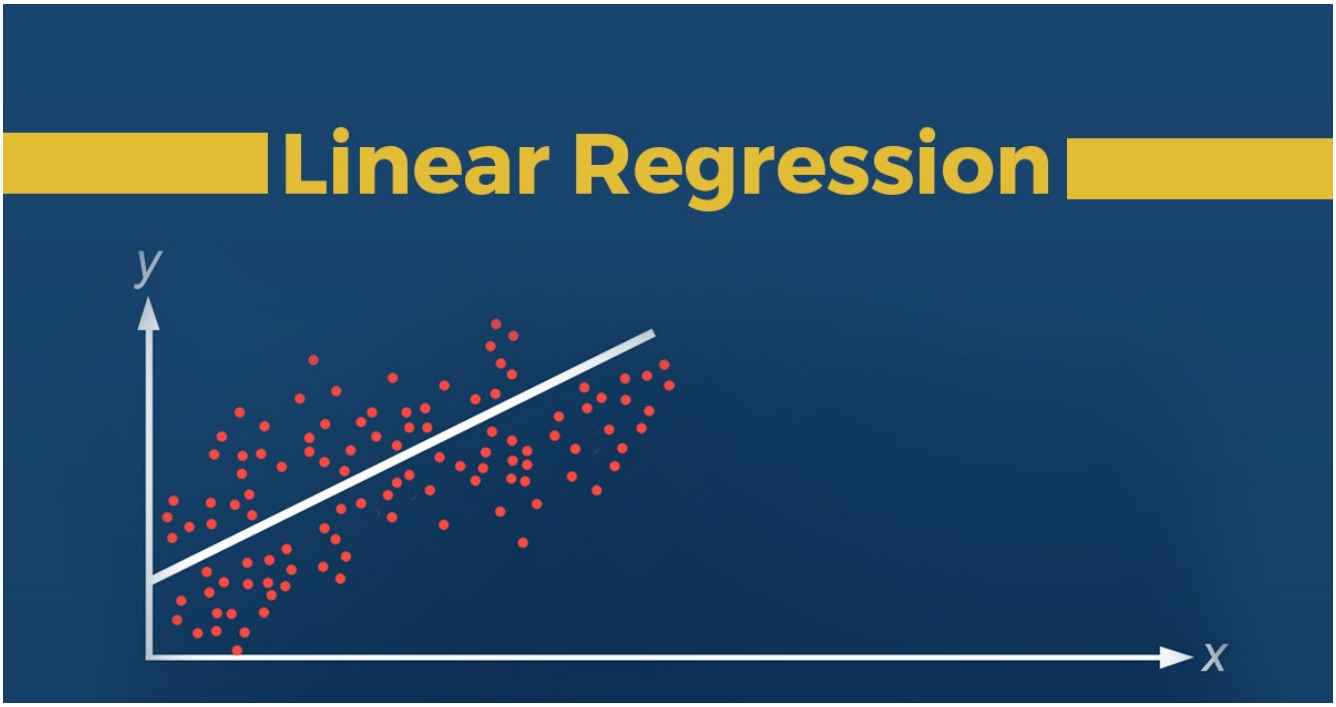


Project: Simple Regression Analysis

University at Buffalo

Pushkaraj Palnitkar

A) Description :

A team of Australian zoologists was formed in July 2020, and tasked with studying the anatomical features of a native species of deer. A sample of 37 adult deer were tranquilized, and measurements recorded on the skull length (measured in millimeters from the back of the skull to the tip of the snout), and neck length (also measured in millimeters).

B) Analysis :

I. Fitting a regression model :

These 37 deer samples are loaded in pandas dataframe, where X is neck length and Y is skull length.

```
with open("df8.txt", "r") as f:
    data = f.read()

x = [float(val) for ind, val in enumerate(data.split()[2:]) if ind%2==0]
y = [float(val) for ind, val in enumerate(data.split()[2:]) if ind%2!=0]

len(x)
```

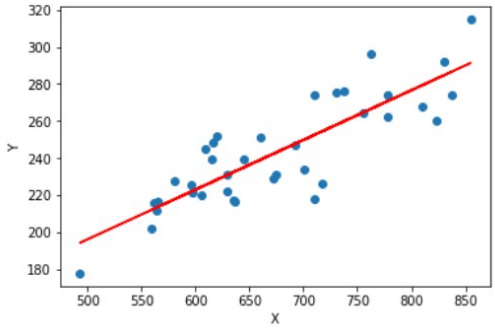
```
37
```

To check linear relationship between X and Y variable we fitted regression line to the data.

```
data = pd.DataFrame({"x":x, "y":y})

# Fit Ordinary Least Squares model
model = ols(formula='y~x',data=data)
result = model.fit()

# Plot
plt.plot(data["x"], result.predict(data["x"]).values, c="r")
plt.scatter(data["x"], data["y"])
plt.xlabel("X")
plt.ylabel("Y");
```



Because, points are well scattered around positive sloped regression line, we can assume linear relationship to begin our analysis.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	62.0771	19.637	3.161	0.003	22.212	101.943
x	0.2682	0.029	9.307	0.000	0.210	0.327

Above table indicates estimated regression parameters.
Note estimated intercept coefficient 62.077 highlights model assumption: In case of zero neck length model assumes 62.077 skull length.

Whereas, estimate of the slope parameter, indicates for each millimeters of increase in the neck length skull length also increases by 0.2682 millimeters.

Estimated R^2:

That means 71.2% of variability in the skull length values can be explained by variability in the neck length values.

```
1 result.rsquared
```

```
0.7122000995576733
```

Unbiased estimate of the parameter σ^2:

Mean Squared Error: 251.909

```
1 result.mse_resid
```

```
251.9092438138481
```

II. Confidence Intervals for β_0 and β_1 :

Obtain confidence intervals for β_0 and β_1. Apply a correction such that the joint confidence level for this set of two intervals is 95%:

$$\alpha' = \alpha/2 = 0.05/2 = 0.025$$

```
1 print("t-statistic:", stats.t.ppf(0.9875,35))
```

```
t-statistic: 2.341969299301038
```

$$\beta_1 = b_1 \pm t_{1-\alpha'/2;n-2}\sqrt{\frac{MSE}{\sum_i (x_i - \bar{x})^2}} = (0.2006, 0.3357)$$

$$\beta_0 = b_0 \pm t_{1-\alpha'/2;n-2}\sqrt{MSE\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right]} = (16.087, 108.067)$$

Hence, we are 95% confident that true b_1 and b_0 together fall in the ranges (0.2006, 0.3357) and (16.087, 108.067) respectively.

III. Hypothesis testing of correlation coefficient (Fisher's Z Transformation) :

It was stated earlier that the team leader guessed that a long neck was associated with also having a long skull. Test (α=0.05) whether there is a linear relationship between neck length and skull length.

$$\textit{Hypothesis Test}$$
$$H_0 : \rho = 0$$
$$H_a : \rho \neq 0$$

Value of ρ :

```
1  np.corrcoef(x,y)[0][1]
```

```
0.8439194864189788
```

Fisher's Z transformation :

$$z = \frac{\sqrt{n-3}}{2}\left[\log\left(\frac{1+R}{1-R}\right) - \log\left(\frac{1+\rho_0}{1-\rho_0}\right)\right] = 3.126$$

```
1  print("p-value:",(1-scipy.stats.norm(0,1).cdf(3.126))*2)
```

```
p-value: 0.0017720155178775343
```

Because, p-value = 0.00177 < 0.05 we can reject the null hypotheses. That means we are 95% confident that neck lengths and skull lengths are correlated with each other.

IV. Model Prediction :

a) Prediction of existing record :

Let's use our fitted regression model to to estimate the skull length of a deer whose neck length measures 600mm.

$$y_{600} = 222.988$$

Let's calculate 90% confidence interval for mean skull length, given that neck length is 600mm.

```
1 print("t-statistics:",stats.t(35).ppf(0.95))
```

```
t-statistics: 1.6895724539637709
```

$$\hat{y}_h \pm t_{1-\alpha/2;n-2}\sqrt{MSE\left[\frac{1}{n}+\frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]} = (217.249, 228.727)$$

Hence, we are 90% confident that the mean skull length when neck length is 600mm is between range 217.249 and 228.727.

b) Prediction of new record :

But, let's now assume one sunny morning, a deer was sighted just outside the team's bunker, and an excited team member tranquilized it. After measuring the neck length as 600mm, the measuring device broke in half. Form a 90% prediction interval for the skull length of this new deer.

$$\hat{y}_h \pm t_{1-\alpha/2;n-2}\sqrt{MSE\left[1+\frac{1}{n}+\frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]} = (195.518, 250.458)$$

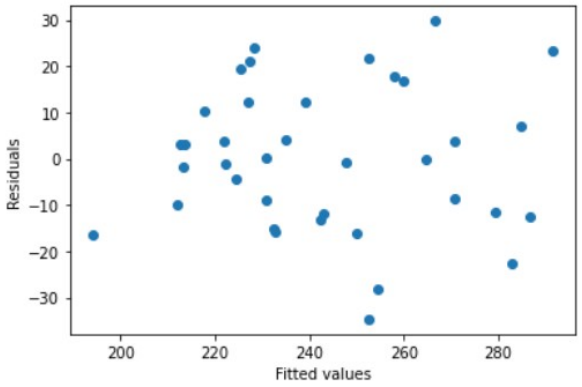
Hence, this time our 90% confidence interval is 195.518, 250.458.

Let's note increase in the confidence interval range when a measurement is taken from outside of the study.

V. Linearity Assumption (Residuals vs fitted values) :

Let's analyze scatter plot of residuals vs fitted values to check linear relationship between two features skull length and neck length:

```
1 plt.scatter(result.fittedvalues, result.resid)
2 plt.xlabel("Fitted values")
3 plt.ylabel("Residuals");
```



From observing above plot we can say that residuals are not the function of fitted values. Because these residuals and fitted values does not exhibit prominent relationship with each other, we can conclude that two features of the model skull length and neck length follows a linear relationship.

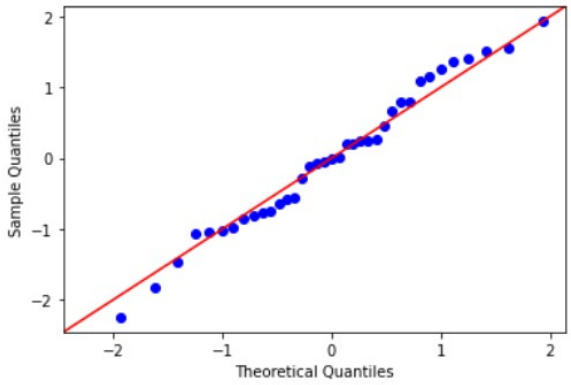
In the above residual plot points are well scattered around the figure, hence, the constant variance assumption of linear regression also holds for the current model.

VI. Normality Assumption :

a) Graphical Method (Q-Q Plot) :

Another linear regression assumption is a normality assumption for the residual values. We will use Q-Q plot to check normality.

```
1 fig = sm.qqplot(result.resid, stats.norm, fit=True, line="45")
```



From the plot above, we can observe sample quantiles (residual quantiles) and theoretical quantiles (normal distribution quantiles) scattered around 45 degrees line. Hence, we can conclude model hold the residual normality assumption.

Note that Q-Q plot is a graphical method for comparing two probability distributions. To have a statistical significance in our normality assumption conclusion we will conduct appropriate hypothesis test.

b) Analytical method (Hypothesis test of normality, Kolmogorov–Smirnov Test) :

Test of residuals for normality:

$$\textit{Hypothesis}$$
$$H_0 : F(x) = G(x)\, i.e., \textit{given data follows normal distribution}$$
$$H_a : F(x) \neq G(x)\, i.e., \textit{given data do not follow normal distribution}$$

```
1 stats.kstest(result.resid.values, "norm", alternative='two-sided')
```

```
KstestResult(statistic=0.45848624069911337, pvalue=1.2388828298945597e-07)
```

Because, p-value < 0.05, we reject the null hypothesis. Hence, we are 95% confident that the residuals are not normally distributed.

This is important to note that QQ plot hints normality assumption but the hypothesis test rejects that assumption with 95% confidence level.

VII. Statistical Power:

Power of the hypothesis test of linear regression model:

$$Hypothesis\ Test:$$
$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

To conduct hypothesis test we need estimated value of the parameter (b_1), that is $\beta_1 = 0.268 \approx 0.3$ and variance of the actual parameter. But because we do not have true variance we assume $var(b_1) = 0.01$.

Power calculation :

$$\delta = \frac{|\beta_1 - 0|}{\sqrt{var(b_1)}} = 3$$

```
1 print("t-statistics:",stats.t(35).ppf(0.975))
```

```
t-statistics: 2.0301079282503425
```

```
1 print("Power:",stats.nct(35,3).cdf(-2.03)+1-stats.nct(35,3).cdf(2.03))
```

```
Power: 0.8306925815064191
```

This means hypothesis test correctly rejects the null hypothesis with probability of 0.8307.

VIII. Obtaining $var(b_1)$ such that statistical power is 90% :

Note the power test calculations are based on our assumption of $var(b_1)$ = 0.01.

To find target $var(b_1)$ we must look back at the power calculation which depends on the value of non-centrality parameter t-test, δ .

As $var(b_1)$ increases the δ decreases, as δ decreases the value of power decreases till it reaches to the value of α. Opposite to that as Var(b_1) decreases the δ increases, as δ increases the value of power increases.

Using this concept we can find target value of $var(b_1)$.

```
beta_1 = 0.3
var_b1 = 0.01
power=0
while power<0.9:
    delta = beta_1/np.sqrt(var_b1)
    t_stats = stats.t(35).ppf(0.975)
    power = stats.nct(35,delta).cdf(-t_stats) + (1-stats.nct(35,delta).cdf(t_stats))
    var_b1-=0.0001

print("Power:",power,"var_b1:",var_b1)
```

```
Power: 0.9033329229654843 var_b1: 0.007900000000000013
```

To achieve 90% statistical power $var(b_1)$ must be less than 0.0079.

C) Conclusion:

1) From the hypothesis test of correlation coefficient we conclude that two variables neck length and skull length are linearly correlated.

2) Fitted linear model explains 71.2% of the variability.

3) Current model holds the linearity assumption but violates the normality assumption.

4) To achieve 90% statistical power true parameter variance var(b_1) must be less than 0.0079.