3.4. The Marketing Plan

Advertising datası için şu 7 soruyu cevaplıyoruz!

1. Reklam bütçesiyle satışlar arasında bir ilişki var mı?

Bu soruya cevap verebilmek için;

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \epsilon$$

TV, radio ve newspaper üzerinden sales'in multiple regression modelini uyguluyoruz ve $H_0: \beta_{TV} = \beta_{radio} = \beta_{newspaper} = 0$ hipotezini test ediyoruz.

F-istatistiği null hipotezi reddedip reddetmeyeceğimizi belirlemek için kullanabileceğimizi öğrenmiştik. Bu üç değişken kullanarak hesapladığımız değerler aşağıdaki gibiydi.

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

F-istatistiği 570 bulmuştuk. Bununla ilişkili p-value değeri oldukça küçük olduğu için null hipotezi reddetmiştik ve bu da reklamlar ile satışlar arasında bir ilişki olduğunu göstermişti.

2. İlişki ne kadar güçlü?

Null hipotezini reddettikten sonra modelin verilere ne ölçüde uyduğunu değerlendirmek için RSE (residual standard error) ve R^2 kullanıyorduk.

- RSE açıklanamayan ($Y = \beta_0 + \beta_1 X + \epsilon$) ϵ'nin standart sapmasının bir tahminiydi. Diğer bir deyişle; Y'nin gerçek regresyon çizgisinden sapacağı ortalama miktardı.

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Advertising datası için RSE = 1,69 çıkmıştı. Y'nin ortalama değeri 14,022 olduğuna göre kabaca %12'lik bir hata vermekte model.

- İkinci olarak R^2, tahmin ediciler kullanılarak modelimizle açıklanabilen varyans Y'deki değişkenlik yüzdesini veriyordu.

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = \frac{\text{Modelin açıkladığı varyans}}{\text{Toplam varyans}}$$

(Toplam varyans; Modelin açıklayamadığı varyans)

Bizim modelimizde $R^2 = 0.897$ çıkmıştı. Yani tahmin ediciler, Y'deki değişkenliğin neredeyse %90'ını açıklıyor.

3. Hangi medya satışlarla ilişkilidir?

Aşağıdaki formülleri kullanarak her tahmin edicinin t-istatistiğini hesaplıyorduk

$$\sigma^2 = Var(\epsilon)$$

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \qquad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Bu formüller simple linear regression için hesaplamıştık. Advertising datası için β_2 de eklenmesi gerekecek ve formüller değişecek. Her bir tahmin edicinin t-istatistiği ile ilişkili p-value'larını aşağıdaki şekilde tabloya yazmıştık

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Her bir tahmin edicinin t-istatistiği ile ilişkili p-value'larına baktığımızda TV ve radyo için p-value'nun düşük olduğunu ama newspaper için düşük olmadığını görüyoruz. Bu, yalnızca TV ve radyonun satışlarla ilgili olduğunu göstermektedir.

4. Her bir ortam ve satış arasındaki ilişki ne kadar büyük?

β_1'in standart hatasının β_1 için güven aralıkları oluşturmak için kullanılabileceğini görmüştük.

β_1 %95 güven aralığında $\beta_1 \pm 2SE(\beta_1)$ aralığında yer alıyordu. Advertising datası için her üç medya bütçesini kullanan bir multiple regresyon modelindeki katsayılar için %95 güven aralığını hesaplamak için yukarıdaki tabloyu kullanabiliriz. Güven aralıkları TV için (0.043, 0.049), radyo için (0.172, 0.206) ve newspaper için (-0.013, 0.011)

TV ve radyo için güven aralıkları dar ve 0'dan uzaktır ve bu medyaların satışlarla ilgili olduğuna dair kanıt sağlar. Ancak newspaper aralığının sıfır olması, TV ve radyo değerleri göz önüne alındığında değişkenin istatistiksel olarak anlamlı olmadığını göstermektedir.

✱ Collinearity'nin çok geniş standart hatalara yol açabileceğini görmüştük. Newspaper ile güven aralığının bu kadar geniş olmasının nedeni collinearity olabilir mi? VIF ile korele olup olmadığına karar veriyorduk

$$VIF = \frac{1}{1 - R^2_{X_j | X_{-j}}}$$

$R^2_{X_j | X_{-j}}$, X_j'nin diğer tüm tahmin ediciler üzerindeki regresyonundan elde edilen değerdi. 1'e yakın olması X_j'yi diğer X_{-j} ile açıklayabildiğimiz, dolayısıyla korelasyon olduğu anlamına gelirdi. R^2'nin 1'e yakın olması ise VIF değerini büyütecektir böyle bir durumda. Genel olarak 5 veya 10'u aşan VIF değeri bize sorunlu bir collinearity'yi haber veriyordu.

Ek not:

VIF 5 ise;

$$1 - R^2 = \frac{1}{5} \Rightarrow R^2 = \frac{4}{5} = 0,8$$

VIF 10 ise;

$$1 - R^2 = \frac{1}{10} \Rightarrow R^2 = \frac{9}{10} = 0,9$$

Advertising datasında TV, radyo ve newspaper için VIF puanları sırasıyla 1.005, 1.145 ve 1.145'tir. Dolayısıyla collinearity olduğuna dair herhangi bir kanıt yoktur.

✱ Her bir ortamın satışlarla ilişkisini ayrı ayrı değerlendirmek için 3 ayrı simple linear regression gerçekleştirebilirdik. TV, radyo ve newspaper için ayrı ayrı yapmış olduğumuz simple linear regressionların istatistiksel değerlerini şöyle gösterebiliriz:

Simple regression of sales on TV

	Coefficients	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	<0.0001
TV	0.0475	0.0027	17.67	<0.0001

Simple regression of sales on radio

	Coefficients	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	<0.0001
radio	0.203	0.020	9.92	<0.0001

Simple regression of sales on newspaper

	Coefficients	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	<0.0001
newspaper	0.055	0.017	3.30	0.00115

Bu tablolarda p-value'lara baktığımızda televizyon ile satış arasında ve radyo ile satış arasında son derece güçlü bir ilişki varken gazete ve satış arasında hafif bir ilişki olduğunu görüyoruz.

5. Gelecekteki satışları ne kadar doğru tahmin edebiliriz?

Yanıtı $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$ formülünü kullanarak tahmin edebiliriz. Bu tahminle ilgili accuracy $Y = f(X) + \epsilon$ bireysel bir yanıtı veya $f(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$ ortalama yanıtı istediğimize bağlıdır.

- $Y = f(X) + \epsilon$ için bir tahmin aralığı kullanırız. Belirli bir şehir için satışları kapsayan tahmin belirsizliğini ölçmek için tahmin aralığı kullanıyorduk ve bu aralık güven aralığına kıyasla çok daha geniş bir aralık oluyordu. Çünkü güven aralığından farklı olarak irreducible hata (ϵ)'yi de hesaba katıyordu.
- $f(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$ için güven aralığı kullanırız. Çok sayıdaki şehirdeki ortalama satışlardaki belirsizliği ölçmek için güven aralığı kullanıyorduk. Bu güven aralığını, aralıkların %95'inin bize gerçek $f(X)$ değerini vereceği şeklinde yorumlamıştık.

6. İlişki lineer mi?

Non-linearityliği belirlemek için residual plots kullanıyorduk. Eğer ilişki lineerse residual plotlar bir pattern göstermiyordu.

TV, radio ve Sales olmak üzere 3 eksen olan grafikte residualları gösterdiğimizde (plazma şeklinde), pozitif residualların (yüzeyin üzerinde görünenler) TV ve radyo bütçelerinin eşit olarak bölündüğü 45 derecelik çizgi boyunca uzanma eğilimindeyken, negatif residuallar bütçelerin daha dengesiz olduğu bu çizgiden uzak durma eğilimindeydiler. Non-linearity ilişkiyi burada gözlemleyebiliyorduk.

Bunun yerine fitted values ($\hat{y}_i$) ve residuals eksenleriyle residual plot grafiği çizilip residualların non-lineer bir pattern izlemesinden de verinin non-lineerliğini gözlemleyebiliyorduk.

Veriye non-linearityliği katmak için polynomial regression'la (X^2) veya $\log X$, $\sqrt{X}$ gibi yöntemlerle predictorı değiştirerek modele non-linearlik katabiliyorduk ve yine lineer regresyon olarak kalıyordu.

$$mpg = \beta_0 + \beta_1 \times horsepower + \beta_2 \times horsepower^2 + \epsilon$$

7. Reklam medya araçları arasında bir sinerji var mı?

Standart lineer regresyon modelleri tahmin ediciler ve yanıt arasında additive bir ilişki olduğunu varsayardı. Fakat bu varsayım her veri için gerçekçi olmayabiliyordu.

Interaction term ekleyip bu değişkenin kat sayısının p-value'sunu bakıyorduk.

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times (radio \times TV) + \epsilon$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Bu terimin p-value'su kaçsa böyle bir ilişki olabileceği anlamına geliyordu. Advertising datası için bu terimi eklediğimizde R^2'nin %90 civarından neredeyse %97'ye çıktığını gördük.