Fen-Edebiyat Fakültesi	**FİNAL SINAV KAĞIDI**	
Adı:	**Dersin Adı: REGRESYON ANALİZİ**	**Not**
Soyadı:	**Dersin Kodu: IST3011**	
Numarası:	**Bölümü: İSTATİSTİK**	
İmzası:	**Sınav Tarihi: 09/02/2021 Saat 13:30-15:30**	

Açıklamalar

1. A4 biçiminde olan **cevap kağıdınızın her birine ad, soyad, okul numarası yazınız ve imza atınız.**
2. Sınav ile ilgili problemleriniz için sınav süresince **fatih.kizilaslan@marmara.edu.tr** e-posta adresinden iletişime geçebilirsiniz.
3. Türkçe haricinde açıklamalar, karalama biçiminde olan yazılar, nereden geldiği belli olmayan tüm ifadeler cevap olarak kabul edilmeyecektir.
   Açıklaması olmayan cevaplar değerlendirilmeyecektir.
4. Cevaplarınızı anlaşılır ve okunabilecek bir biçimde sisteme yükleyiniz.
5. **Bu sınava katılan her öğrenci bu kuralları ve önceden ilan edilmiş tüm kuralları kabul etmiş olarak değerlendirilecektir.**

SINAV İLE İLGİLİ AÇIKLAMALAR

Cevaplarınızı R Markdown kullanarak oluşturunuz. Yazmanız gereken matematiksel ifadeleri soru numarasını yazarak A4 kağıdına yazabilirsiniz. Oluşturduğunuz R Markdown ve A4 kağıdındaki çözümlerinizi birleştirerek PDF formatında sisteme yükleyiniz.

Sınav sonunda ile ilgili **R Markdown kodunuzu**n adını **"isim_soyisim"** olarak kaydederek **e-posta** ile **fatih.kizilaslan@marmara.edu.tr** adresine gönderiniz.

Her bir soruyu R Markdown'da sıra numarasını yazarak cevaplayınız.

Kaggle'da "https://www.kaggle.com/mirichoi0218/insurance?select=insurance.csv" adresinde yer alan (ayrıca sınavdan bir kaç dakika önce BYS'de bulunan **e-posta adreslerinize ve UES sistemi üzerinden gönderdiğim** "insurance.csv") kişilerin sağlık sigortaları tarafından ödenen sağlık harcamaları ve bunu etkileyen bazı değişkenleri içeren veriyi kullanarak aşağıdaki soruları cevaplayınız.

Bu veri toplam 1338 gözlem ve 7 değişkenden oluşmaktadır. Değişkenler aşağıdaki gibidir.

age	**sex**	**bmi**	**children**	**smoker**	**region**	**charges**
yaş	cinsiyet	vücut kitle indeksi	çocuk sayısı	sigara içme durumu	yaşanılan bölge	sigortanın ödediği miktar
	2 kategori			2 kategori	4 kategori	

Bu analiz için anlamlılık düzeyi $\alpha = 0.05$ olarak alınacaktır. Sadece sorularda sizden istenilen soruları açık ve en kısa bir biçimde açıklayınız.

Yapacak olduğunuz gereksiz analiz, sonuç, grafik vs gibi işlemler sınav sonucunun açıklanmasını uzatacaktır. Bu durum Bütünleme sınavınıza daha az hazırlanma sürenizin oluşmasına sebep olacaktır.

SORULAR

1. **(5 puan)** Okul numaranızın **6. basamağındaki rakam** a ve **son iki basamağındaki sayı** b olarak alarak **insurance.csv** verisinin ilk $500+[10*(a+b)]$ gözlemini kullanarak "**my_data**" adında data.frame oluşturunuz.

   Örneğin, okul numaranız 121507085 ise $a = 7$ ve $b = 75$ olmak üzere **insurance.csv** verisinin ilk $500 + [10 * (75 + 7)] = 1320$ gözlemi ile **my_data** oluşturulur.

2. **(6 puan) sex, smoker** ve **region** değişkenlerini **gösterge (dummy) değişken** olarak tanımlayınız.

3. **(9 puan) charges** bağımlı değişken ve **age, sex, bmi, smoker, region** bağımsız değişkenler olmak üzere çoklu doğrusal regresyon modelini matematiksel olarak ifade ediniz. (Değişken tanımlamalarınızı $y, x_1, x_2, ...$ kullanarak yazınız.)

4. **(10 puan)** 3. soruda yazdığınız matematiksel modeli **Model_1** olarak tanımlayarak çoklu doğrusal regresyon modelini oluşturunuz. Model_1'in anlamlılığı için gerekli hipotezleri yazınız. Model anlamlı mıdır? Açıklayınız. R^2 ve R^2_{adj} değerlerini yorumlayınız.

5. **(5 puan)** Model_1'den **region** değişkenini çıkararak **Model_2** oluşturunuz. Model_2 anlamlı mıdır? Açıklayınız

6. **(10 puan)** Model_2'deki değişkenlerin anlamlılıklarını kısmi t testlerine göre yorumlayınız.

7. **(20 puan)** Model_2'nin sonuçlarına göre **sex** ve **smoker** değişkenlerine göre oluşabilecek tüm regresyon denklemlerini yazınız. Herhangi 2 tanesini karşılaştırarak yorumlayınız.

8. **(10 puan)** Model_2 için R^2 ve R^2_{adj} değerlerini yorumlayınız. Model_1 ile karşılaştırarak yorumlayınız.

9. **(10 puan) region** değişkeninin anlamlığını kısmi F testi ile test ediniz (sadece R programını kullanarak). Sonuçlarını açıklayınız. Ayrıca, yaptığınız bu testin hipotezlerini yazınız.

10. **(5 puan)** Kendiniz için **age, sex, bmi, smoker** değerlerini oluşturarak bu değerlere karşılık gelen **charges** yanıt değişkeninin için tahmin değerini bulunuz.

    Not: bmi=$\dfrac{\text{ağırlık(kg)}}{(\text{metre cinsinden boy uzunluğunuz})^2}$

11. **(10 puan)** 10. soruda bulduğunuz bağımsız değişkenlerin değerleri için %95 güven düzeyinde tahmin aralığını bulunuz. Sonucu yorumlayınız.

BAŞARILAR **Doç. Dr. Fatih KIZILASLAN**