Erneut werden die Schweizer Theaterdaten vom Blatt 10 betrachtet:

id	alter	geschlecht	gehalt	kultur	theaterheute	theatergestern	gehalt_kat
1	31.00	weiblich	90.50	181.00	104.00	150.00	>90
2	54.00	männlich	73.00	234.00	116.00	140.00	70-80
3	56.00	weiblich	74.30	289.00	276.00	125.00	70-80

Aufgabe 1: Es werden die Theaterausgaben gestern und heute betrachtet.

a) Sie erhalten folgenden R-Output:

```
        Pearson's product-moment correlation

data:  theater$theaterheute and theater$theatergestern
t = 5.378, df = 697, p-value = 1.029e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1273372 0.2697709
sample estimates:
      cor
0.1996082
```

(i) Welche Größe wurde hier berechnet?

(ii) Wie Sie am Output erkennen können, wurde auch ein Test durchgeführt. Wie lauten dessen Hypothesen? Wie entscheidet der Test?

b) Sie berechnen nun, anstatt des obigen Tests, eine lineare Einfachregression:

```
Call:
lm(formula = theaterheute ~ theatergestern, data = theater)

Residuals:
    Min      1Q  Median      3Q     Max
-141.59  -53.48  -19.76   36.40  314.24

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     47.3709    17.3970   2.723  0.00663 **
theatergestern   0.6759     0.1257   5.378 1.03e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 75.03 on 697 degrees of freedom
Multiple R-squared:  0.03984,        Adjusted R-squared:  0.03847
F-statistic: 28.92 on 1 and 697 DF,  p-value: 1.029e-07
```

(i) Was ist die Zielgröße?

(ii) Welche Parallelen sehen Sie zu dem obigen Test?

Aufgabe 2: In dieser Aufgabe wird erneut der Zusammenhang zwischen Geschlecht und den Kulturausgaben untersucht.
Wie bereits in Aufgabenblatt 10 wird ein doppelter t-Test durchgeführt:

```
Welch Two Sample t-test

data:  kultur by geschlecht
t = 1.3018, df = 667.43, p-value = 0.1934
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.602222 12.841554
sample estimates:
mean in group männlich mean in group weiblich
              222.7120               217.5923
```

Dieser Test, der sog. „Welch-Test“, nimmt **nicht** an, dass die theoretischen Varianzen in den beiden Gruppen gleich sind. Würde man dies jedoch annehmen, so erhielte man folgendes Testergebnis:

```
Two Sample t-test

data:  kultur by geschlecht
t = 1.2983, df = 697, p-value = 0.1946
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.622518 12.861851
sample estimates:
mean in group männlich mean in group weiblich
              222.7120               217.5923
```

Wie bereits in Aufgabe 1, kann auch hier anstatt des einfachen Tests eine lineare Einfachregression durchgeführt werden:

```
Call:
lm(formula = kultur ~ geschlecht, data = theater)

Residuals:
    Min      1Q  Median      3Q     Max
-158.71  -34.65   -3.59   29.41  400.41

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         222.712      2.945  75.612   <2e-16 ***
geschlechtweiblich   -5.120      3.943  -1.298    0.195
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.78 on 697 degrees of freedom
Multiple R-squared:  0.002413,  Adjusted R-squared:  0.0009813
F-statistic: 1.686 on 1 and 697 DF,  p-value: 0.1946
```

a) Welche Zusammenhänge erkennen Sie zwischen den beiden durchgeführten Tests und der linearen Einfachregression?

b) Welche Annahme steckt hinter dem p-Wert der Regressionskoeffizienten?

Aufgabe 3: Abschließend wird auch der Zusammenhang zwischen Gehalt (kategorial) und Kulturausgaben untersucht.

a) Zunächst führen Sie eine ANOVA (**An**alysis **o**f **Va**riance) durch:

```
Analysis of Variance Table

Response: kultur
            Df  Sum Sq Mean Sq F value   Pr(>F)
gehalt_kat   4   37744  9435.9  3.5681 0.006838 **
Residuals  694 1835305  2644.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(i) Was wird mit einer ANOVA überprüft?
Stellen Sie einen Zusammenhang zum t-Test her!

(ii) Wie lauten die Hypothesen und die Testentscheidung?

(iii) Was können Sie diesem Output für Informationen entnehmen?

b) Nun untersuchen Sie den Zusammenhang mittels eines Regressionsmodells.

```
Call:
lm(formula = kultur ~ gehalt_kat, data = theater)

Residuals:
    Min      1Q  Median      3Q     Max
-151.42  -33.42   -4.42   29.34  409.34

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      208.655      4.775  43.700  < 2e-16 ***
gehalt_kat60-70    6.769      5.823   1.162  0.24545
gehalt_kat70-80   15.760      6.030   2.614  0.00915 **
gehalt_kat80-90   21.899      7.393   2.962  0.00316 **
gehalt_kat>90     19.808      7.891   2.510  0.01230 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.43 on 694 degrees of freedom
Multiple R-squared:  0.02015,        Adjusted R-squared:  0.0145
F-statistic: 3.568 on 4 and 694 DF,  p-value: 0.006838
```

Welche Zusammenhänge zur ANOVA erkennen Sie?

Aufgabe 4: Abschließende Fragen:

a) Wo sehen Sie Unterschiede zwischen der multiplen Regression und einfachen Mittelwertsvergleichen?

b) Wo sehen Sie Vorteile/Nachteile?