The Story: Billionaires 1992.

Fortune magazine publishes a list of the world's billionaires each year. The data set used in this exam contains 175 individuals (a subset of the whole list) from the 1992 list. Their `wealth`, `age`, and geographic location (Asia, Europe, United States), i.e., `region`, are included in a data set named `Billionaire`, where the first 37 observations are billionaires from Asia, then the next 76 are from Europe, and the remaining 62 are from United States.

```
> table(Billionaire$region)
 A  E  U
37 76 62

> Billionaire
    wealth age region
1     14.0  64      A
2      4.9  62      A
......
37     1.0  69      A
38    11.7  72      E
39     7.2  66      E
......
113    1.0  59      E
114   24.0  88      U
......
175    1.0  63      U
```

Problem 1 : Fit a one-way ANOVA model with `wealth` as the response and `region` as the predictor.

1. Check for a good transformation on the response. Based on the box-cox plot, explain which of the following transformation is appropriate:

$$1)\ \log y, \quad 2)\ y^{-1}, \quad 3)\ \text{no transformation is needed.}$$

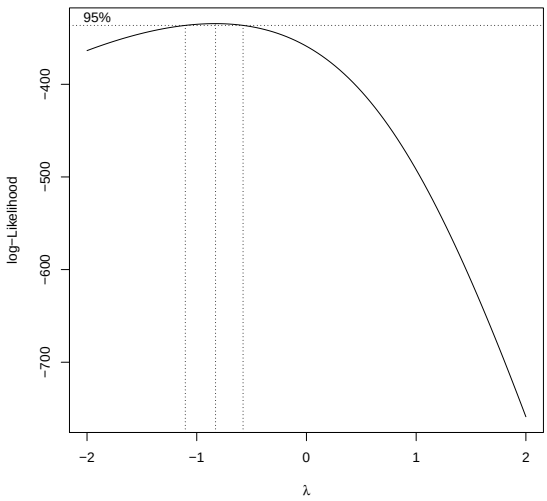

Figure 1: Box-cox transformation

Define `newwealth` as the transformed response, and carry out the remaining analysis using `newwealth`.

```
> g0 = lm(newwealth ~ region)
> summary(g0)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.50756
regionE      0.07519
regionU      0.04167

> anova(g0)
           Df  Sum Sq  Mean Sq F value Pr(>F)
region    xxx  0.1439   xxxx    xxxx   0.319
Residuals xxx 10.7561

> mean(newwealth[region=='A'])
[1] 0.5076
> mean(newwealth[region=='E'])
[1] 0.5827
> mean(newwealth[region=='U'])
[1] 0.5492
```

2. We want to test whether there are significant differences among regions. What's the value of your test statistic? What's its distribution under the null? What's your conclusion?

3. Consider the following pairwise differences:

$$\begin{aligned} \gamma_1 &: \text{Asia vs. United States} \\ \gamma_2 &: \text{Europe vs. United States} \\ \gamma_3 &: \text{Asia vs. Europe} \end{aligned}$$

   Calculate the LS estimates for γ_j $(j = 1, 2, 3)$, as well as their standard deviations.

4. Suppose we want to construct simultaneous 95% CIs for γ_j $(j = 1, 2, 3)$ using Bonferroni correction. Let t_v^α denote the critical value for the corresponding student T distribution. What are the values for v and α?

5. Suppose we want to construct simultaneous 95% CIs for γ_j $(j = 1, 2, 3)$ using Tukey's method. Let the $q_{m,v}^\alpha$ denote the critical value for the corresponding studentized range distribution. What are the values for m, v and α?

6. Suppose we want to construct simultaneous 95% CIs for γ_j $(j = 1, 2, 3)$ using Scheffé's method. Let the F_{v_1,v_2}^α denote the critical value for the corresponding F distribution. What are the values for v_1, v_2 and α?

Problem 2 : Fit a simple regression model with `newwealth` as the response and `age` as the predictor. Obtain a test for lack of fit of this straight-line model. Based on the `R` output below, find the value of your test statistic. What's its distribution under H_0?

```
> summary(lm(newwealth ~ age))$sigma
[1] 0.2509

 # Among the 175 obs, there are only 50 unique values for age.
> length(unique(age))
[1] 50
> summary(lm(newwealth ~ as.factor(age)))$sigma
[1] 0.2483
```

Problem 3 : Fit an ANCOVA model for `newwealth` with the continuous explanatory variable `age` and the discrete explanatory variable `region`. Here assume it is still appropriate to apply the same transformation on `wealth` as what we did for the one-way ANOVA model.

```
> g = lm(newwealth ~ age * region)
> summary(g)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4105
age           0.0015
regionE       0.1875
regionU       0.0552
age:regionE  -0.0018
age:regionU  -0.0002

Residual standard error: 0.252 on xxx degrees of freedom

> model.matrix(g)
    (Intercept) age regionE regionU age:regionE age:regionU
1             1  64       0       0           0           0
2             1  62       0       0           0           0
......
37            1  69       0       0           0           0
38            1  72       1       0          72           0
39            1  66       1       0          66           0
......
113           1  59       1       0          59           0
114           1  88       0       1           0          88
115           1  63       0       1           0          63
......
175           1  63       0       1           0          63
```

1. Model `g` implies a linear model: `newealth` $= a + b \cdot$ `age` for billionaires from Europe. What's a and what's b?

2. The 82nd observation has the largest leverage, who is a 7-year-old billionaire from Europe. Is the following statement true? Explain why. "Since the 82nd observation has the largest leverage, it must be an outlier."

   ```
   > lev = influence(g)$hat
   > sort(lev, decreasing=TRUE)[1:3]  # top 3 leverages
       82    123      7
   0.2139 0.2128 0.1707
   ```

```
> Billionaire[82, ]
   wealth age region
82    xxx   7      E

> g$res[82]
0.07037
```

3. Based on the normal assumption for linear regression models, the residual of the 82nd observation follows a normal distribution. What's the mean and what's the variance? (If your answer depends on any unknown parameter of the linear regression model, use the corresponding estimate).

4. To evaluate the influence of the 82nd observation, we can refit the model based on the 174 observations (excluding the 82nd observation). Let a denote the prediction of `newwealth` for a 7-yr-old billionaire from Europe based on this new model. What's the difference between the observed `newwealth` (for the 82nd observation) and a?

5. To test whether the 82nd observation is an outlier, we need to compare its jackknife residual to a critical value t_v^α. Suppose we want a level 5% test.

   a) If we do not know which observation is an outlier before analyzing the data, how should we set the value for α and the value for v?

   b) It is reasonable to suspect that the 82nd sample is an outlier due to his young age before we run the ANCOVA model. Then how should we set the value for α and the value for v?

6. Model `g` implies three different linear models:

$$\texttt{newwealth} = \alpha_k + \beta_k \cdot \texttt{age}, \quad k = 1, 2, 3,$$

   for the three regions. Based on the `R` output below, can we reduce `g` to one of the following models?

   I) α_k's are different, but the slopes are the same, $\beta_1 = \beta_2 = \beta_3$, i.e., the three lines are parallel.

   II) $\alpha_1 = \alpha_2 = \alpha_3$ and $\beta_1 = \beta_2 = \beta_3$, i.e., one line for all three regions.

III) $\alpha_1 = \alpha_2 = \alpha_3$ and $\beta_1 = \beta_2 = \beta_3 = 0$, i.e., an intercept only model.

```
> anova(lm(newwealth ~ age * region))

            Df   F value Pr(>F)
age         xxx    0.1002 0.7520
region      xxx    1.1281 0.3261
age:region xxx    0.1422 0.8675
Residuals  xxx

> anova(lm(newwealth ~ region * age))

            Df   F value Pr(>F)
region      xxx    1.1326 0.3246
age         xxx    0.0912 0.7630
region:age xxx    0.1422 0.8675
Residuals  xxx
```

Problem 1 : Fit a one-way ANOVA model with `wealth` as the response and `region` as the predictor.

1. The box-cox transformation defines a family of power transformations on the response y, namely, $(y^\lambda - 1)/\lambda$. The log-transformation corresponds to $\lambda = 0$, and no transformation corresponds to $\lambda = 1$. The plot indicates that the 95% CI for λ contains -1, but not 0 nor 1. So the transformation $1/y$ is appropriate.

2. We can use the F-test. The value for the F-stat is

$$\frac{0.1439/2}{10.7561/172} = 1.1505.$$

   It follows $F_{2,172}$. The corresponding p-value is 0.319, so no strong evidence supporting the one-way ANOVA model, that is, there is no significant difference of the response variable `newwealth` among different regions.

3. $\hat{\sigma} = \sqrt{10.7561/172} = 0.06254$.

$$\begin{aligned} \hat{\gamma}_1 &= -0.04167, & se(\hat{\gamma}_1) &= \hat{\sigma}\sqrt{1/37 + 1/62}; \\ \hat{\gamma}_2 &= 0.03352, & se(\hat{\gamma}_2) &= \hat{\sigma}\sqrt{1/76 + 1/62}; \\ \hat{\gamma}_3 &= -0.07519, & se(\hat{\gamma}_3) &= \hat{\sigma}\sqrt{1/37 + 1/76}. \end{aligned}$$

4. $v = 172$ and $\alpha = 5\%/6$.

5. $m = 3, v = 172$, and $\alpha = 5\%$.

6. $v_1 = 2, v_2 = 172$, and $\alpha = 5\%$.

Problem 2 : Lack-of-fit test. The test statistic is

$$\frac{(0.2509^2 \times 173 - 0.2483^2 \times 125)/(173 - 125)}{0.2483^2} = 1.076,$$

which follows $F_{48,125}$.

Problem 3 : Fit an ANCOVA model.

1. $a = 0.4105 + 0.1875$ and $b = 0.0015 - 0.0018$.

2. The statement is NOT true.

   High leverage samples are samples that are far away from others (in the feature space), and we judge whether a sample has a high leverage only based on the design matrix $\mathbf{X}$.

   Outliers are samples with a large leave-one-out residual, i.e., outliers are samples that do not fit a regression model determined by the other samples; to determine whether a sample is an outlier, we have to look at both the design matrix $\mathbf{X}$ and the response vector $\mathbf{y}$.

   So these two concepts are different, and a high leverage sample is not necessarily an outlier.

3. The residual (for the 82nd sample) $\sim \mathsf{N}(0, \hat{\sigma}^2(1 - 0.2139))$ where $\hat{\sigma} = 0.252$.

4. The difference (i.e., leave-one-out residual) is equal to $0.07037/(1 - 0.2139) = 0.08952$.

5. (Note the jackknife residual is the normalized version of the leave-one-out residual.)

   a) $\alpha = 5\%/(2 \times 175)$ and $v = (174 - 6) = 168$.

   b) $\alpha = 5\%/2$ and $v = 168$.

6. The intercept only model.

STAT 425 *Sample* Exam #2

November 18, 2014
9:30 – 10:50am

This exam is closed book and closed notes. You are allowed to use a calculator and a formulae sheet (double-sided), but materials may not be shared. There are three (blank) work-sheets at the end. Please let me know if you need more.

- You are required to provide complete justification for each problem in order to get full credit.
- Please give all numerical answers to at least **three** correct digits or as exact fractions reduced to lowest terms.
- Write your solutions as clearly as possible and make sure it's easy to find your answers (circle them if necessary).

If you find a question confusing please **ask me** to clarify it.

Print Name: _____

1.	/ xx
2.	/ xx
3.	/ xx
Total:	/ xx

The Story: Billionaires 1992.

Fortune magazine publishes a list of the world's billionaires each year. The data set used in this exam contains 175 individuals (a subset of the whole list) from the 1992 list. Their `wealth`, `age`, and geographic location (Asia, Europe, United States), i.e., `region`, are included in a data set named `Billionaire`, where the first 37 observations are billionaires from Asia, then the next 76 are from Europe, and the remaining 62 are from United States.

```
> table(Billionaire$region)
 A  E  U
37 76 62

> Billionaire
    wealth age region
1     14.0  64      A
2      4.9  62      A
......
37     1.0  69      A
38    11.7  72      E
39     7.2  66      E
......
113    1.0  59      E
114   24.0  88      U
......
175    1.0  63      U
```

Problem 1 : Fit a one-way ANOVA model with `wealth` as the response and `region` as the predictor.

1. Check for a good transformation on the response. Based on the box-cox plot, explain which of the following transformation(s) is(are) appropriate:

   1) $\log y$, 2) $1 - 1/y$, 3) $y^{-2/3}$, 4) no transformation is needed.

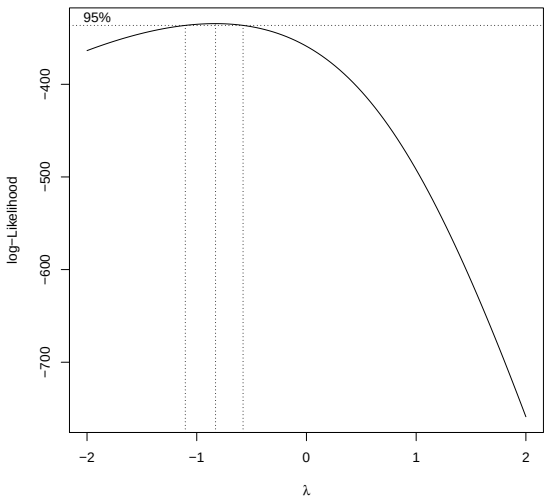

Figure 1: Box-cox transformation

Define `newwealth` as the transformed response, and carry out the remaining analysis using `newwealth`.

```
> g0 = lm(newwealth ~ region)
> summary(g0)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.50756
regionE      0.07519
regionU      0.04167

> anova(g0)
           Df  Sum Sq  Mean Sq F value Pr(>F)
region    xxx  0.1439   xxxx    xxxx   0.319
Residuals xxx 10.7561

> mean(newwealth[region=='A'])
[1] 0.5076
> mean(newwealth[region=='E'])
[1] 0.5827
> mean(newwealth[region=='U'])
[1] 0.5492
```

2. We want to test whether there are significant differences among regions. What's the value of your test statistic? What's its distribution under the null? What's your conclusion?

3. Consider the following pairwise differences:

$$\begin{aligned} \gamma_1 &: \text{Asia vs. United States} \\ \gamma_2 &: \text{Europe vs. United States} \\ \gamma_3 &: \text{Asia vs. Europe} \end{aligned}$$

   Calculate the LS estimates for γ_j $(j = 1, 2, 3)$, as well as their standard deviations.

4. Suppose we want to construct simultaneous 95% CIs for γ_j $(j = 1, 2, 3)$ using Bonferroni correction. Let t_v^α denote the critical value for the corresponding student T distribution. What are the values for v and α?

5. Suppose we want to construct simultaneous 95% CIs for γ_j $(j = 1, 2, 3)$ using Tukey's method. Let the $q_{m,v}^\alpha$ denote the critical value for the corresponding studentized range distribution. What are the values for m, v and α?

6. Suppose we want to construct simultaneous 95% CIs for γ_j $(j = 1, 2, 3)$ using Scheffé's method. Let the F_{v_1,v_2}^α denote the critical value for the corresponding F distribution. What are the values for v_1, v_2 and α?

Problem 2 : Fit a simple regression model with `newwealth` as the response and `age` as the predictor.

1. Obtain a test for lack of fit of this straight-line model. Based on the `R` output below, find the value of your test statistic. What's its distribution under H_0?

```
> summary(lm(newwealth ~ age))$sigma
[1] 0.2509

 # Among the 175 obs, there are only 50 unique values for age.
> length(unique(age))
[1] 50
> summary(lm(newwealth ~ as.factor(age)))$sigma
[1] 0.2483
```

2. Is the following statement true? Explain why. "If the lack of fit test is not significant, then any nonlinear polynomial model is worse than the linear model for this data set." Here "worse" means that if we run an F-test to compare a nonlinear polynomial model with the linear model, then we cannot reject the linear model.

Problem 3 : Fit a linear model for `newwealth` with the continuous explanatory variable `age` and the discrete explanatory variable `region`. Here assume it is still appropriate to apply the same transformation on `wealth` as what we did for the one-way ANOVA model.

```
> g = lm(newwealth ~ age * region)
> summary(g)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4105
age           0.0015
regionE       0.1875
regionU       0.0552
age:regionE  -0.0018
age:regionU  -0.0002

Residual standard error: 0.252 on xxx degrees of freedom

> model.matrix(g)
    (Intercept) age regionE regionU age:regionE age:regionU
1             1  64       0       0           0           0
2             1  62       0       0           0           0
......
37            1  69       0       0           0           0
38            1  72       1       0          72           0
39            1  66       1       0          66           0
......
113           1  59       1       0          59           0
114           1  88       0       1           0          88
115           1  63       0       1           0          63
......
175           1  63       0       1           0          63
```

1. The 82nd observation has the largest leverage, who is a 7-year-old billionaire from Europe. Is the following statement true? Explain why. "Since the 82nd observation has the largest leverage, it must be an outlier."

```
> lev = influence(g)$hat
> sort(lev, decreasing=TRUE)[1:3]  # top 3 leverages
    82    123      7
0.2139 0.2128 0.1707

> Billionaire[82, ]
   wealth age region
82    xxx   7      E

> g$res[82]
0.07037
```

2. Based on the normal assumption for linear regression models, the residual of the 82nd observation follows a normal distribution. What's the mean and what's the variance? (If your answer depends on any unknown parameter of the linear regression model, use the corresponding estimate).

3. To evaluate the influence of the 82nd observation, we can refit the model based on the 174 observations (excluding the 82nd observation). Let a denote the prediction of `newwealth` for a 7-yr-old billionaire from Europe based on this new model. What's the difference between the observed `newwealth` (for the 82nd observation) and a?

4. To test whether the 82nd observation is an outlier, we need to compare its studentized residual to a critical value t_v^α. Suppose we want a level 5% test.

   a) If we do not know which observation is an outlier before analyzing the data, how should we set the value for α and the value for v?

   b) It is reasonable to suspect that the 82nd sample is an outlier due to his young age before we fit the model. Then how should we set the value for α and the value for v?

Problem 1 : Fit a one-way ANOVA model with `wealth` as the response and `region` as the predictor.

1. The box-cox transformation defines a family of power transformations on the response y, namely, $(y^\lambda - 1)/\lambda$. The log-transformation corresponds to $\lambda = 0$, and no transformation corresponds to $\lambda = 1$. The plot indicates that the 95% CI for λ contains -1 and $-2/3$, but not 0 nor 1. So (2) and (3) are appropriate. Note that once we pick λ ($\lambda \neq 1$), we could define the new y by y^λ or $(y^\lambda - 1)/\lambda$ or any linear transformation of y^λ.

2. We can use the F-test. The value for the F-stat is

$$\frac{0.1439/2}{10.7561/172} = 1.1505.$$

   It follows $F_{2,172}$. The corresponding p-value is 0.319, so no strong evidence supporting the one-way ANOVA model, that is, there is no significant difference of the response variable `newwealth` among different regions.

3. $\hat{\sigma} = \sqrt{10.7561/172} = 0.06254$.

$$\begin{aligned} \hat{\gamma}_1 &= -0.04167, & se(\hat{\gamma}_1) &= \hat{\sigma}\sqrt{1/37 + 1/62}; \\ \hat{\gamma}_2 &= 0.03352, & se(\hat{\gamma}_2) &= \hat{\sigma}\sqrt{1/76 + 1/62}; \\ \hat{\gamma}_3 &= -0.07519, & se(\hat{\gamma}_3) &= \hat{\sigma}\sqrt{1/37 + 1/76}. \end{aligned}$$

4. $v = 172$ and $\alpha = 5\%/6$.

5. $m = 3, v = 172$, and $\alpha = 5\%$.

6. $v_1 = 2, v_2 = 172$, and $\alpha = 5\%$.

Problem 2 : Lack-of-fit test.

1. The test statistic is

$$\frac{(0.2509^2 \times 173 - 0.2483^2 \times 125)/(173 - 125)}{0.2483^2} = 1.076,$$

   which follows $F_{48,125}$.

2. The statement is wrong. The lack-of-fit test compares the linear model with the model with the largest df which we could fit on this data set. Failing to reject the linear model does not imply that there does not exist a nonlinear model with a modest df which is better than the linear model.

Problem 3 :

1. The statement is NOT true.

   High leverage samples are samples that are far away from others (in the feature space), and we judge whether a sample has a high leverage only based on the design matrix $\mathbf{X}$.

   Outliers are samples with a large leave-one-out residual, i.e., outliers are samples that do not fit a regression model determined by the other samples; to determine whether a sample is an outlier, we have to look at both the design matrix $\mathbf{X}$ and the response vector $\mathbf{y}$.

   So these two concepts are different, and a high leverage sample is not necessarily an outlier.

2. The residual (for the 82nd sample) $\sim \mathsf{N}(0, \hat{\sigma}^2(1 - 0.2139))$ where $\hat{\sigma} = 0.252$.

3. The difference (i.e., leave-one-out residual) is equal to $0.07037/(1 - 0.2139) = 0.08952$.

4. (Note the studentized residual is the normalized version of the leave-one-out residual.)

   a) $\alpha = 5\%/(2 \times 175)$ and $v = (174 - 6) = 168$.

   b) $\alpha = 5\%/2$ and $v = 168$.