MSG500 Final 2016-08-25

Examiner: Rebecka Jörnsten, 0760-491949

Remember: To pass this course you also have to hand in a final project to the examiner.
Open book, open notes but no calculators or computers allowed

Make sure to give detailed and specific answers. Avoid yes/no answers. You should also provide a motivation. Good Luck!

Question 1(35=5+5+5+5+5+5+5)

The NHANES data set is a very large public health data set. Here you will look at a small subset of subjects and variables recorded. Specifically, you will look at the cholesterol level of 1000 subjects. Other variables recorded are; gender, age, HDL (the "good" cholesterol), DBP, SBP (dpb is a more reliable marker for cardiovascular risk than systolic bp), weight (wt), height (ht) and body mass index (bmi).
a) Below I summarize the fit of a regression model including all variables. Comment on the appropriateness of the fit, the meaning of the model, if you think variables have to be transformed etc. Make sure you state how you come to your finding. Discuss other information you might want to have access to to say more about the fit and model.

Solution: From Figure 1, the scatter plots of cholesterol vs some variables exhibit a weak relationship. In plots vs hdl and sbp there is an indication of non-constant error variance - probably we should transform cholesterol (and some of the xs) with e.g. log or square root. In Figure 2 I see very weak relationships between chol and the other variables. I also see that some strong correlations between the xs (e.g. wt and ht).
The residual diagnostics indicate that the error distribution is skewed (long right-tail) - we really should transform y (and some xs). There is no trend in the residuals to speak of and outliers are not extreme (and will have limited impact in such a big data set).
We need to see the residual plots vs x variables to check for other trends or nonconstant error variance. Coplots to check for interactions also (big data - we can afford interactions). What's the meaning of the model? Looking at the model summary we see that hdl and dbp are significant. Taken at face value the model states that hdl is positively related to cholesterol and dpb is as well. However, the R-squares indicates that only 10% of the variability in cholesterol is explained by the included variables (less than that if we only use the significant variables). There is not much explantatory power in this data set about cholesterol.

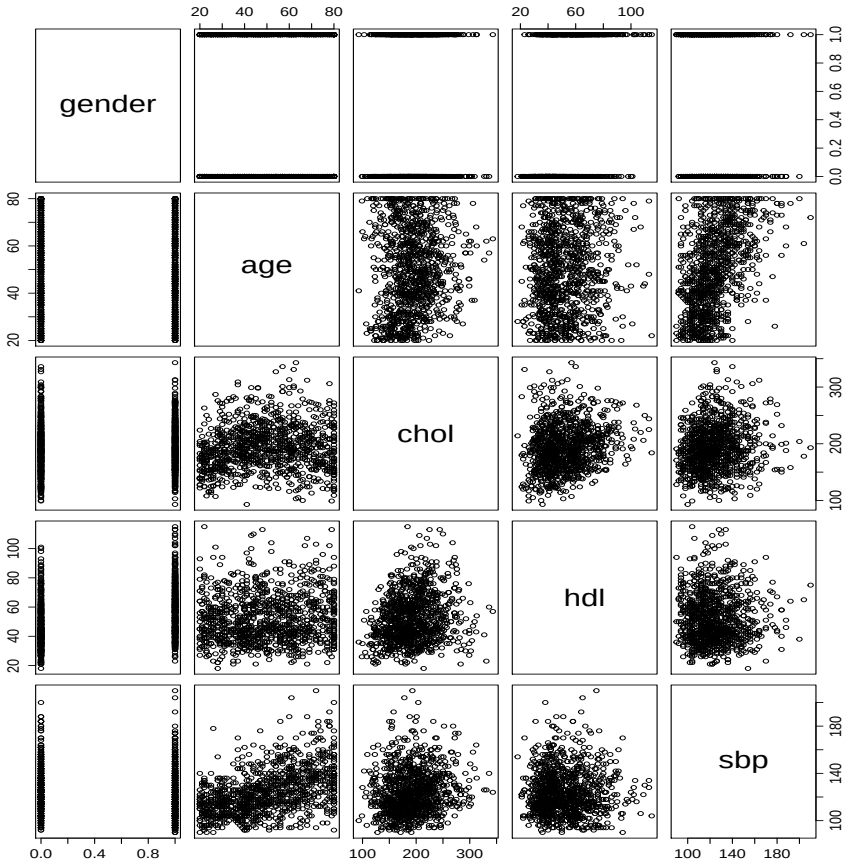

Figure 1: Scatter plots

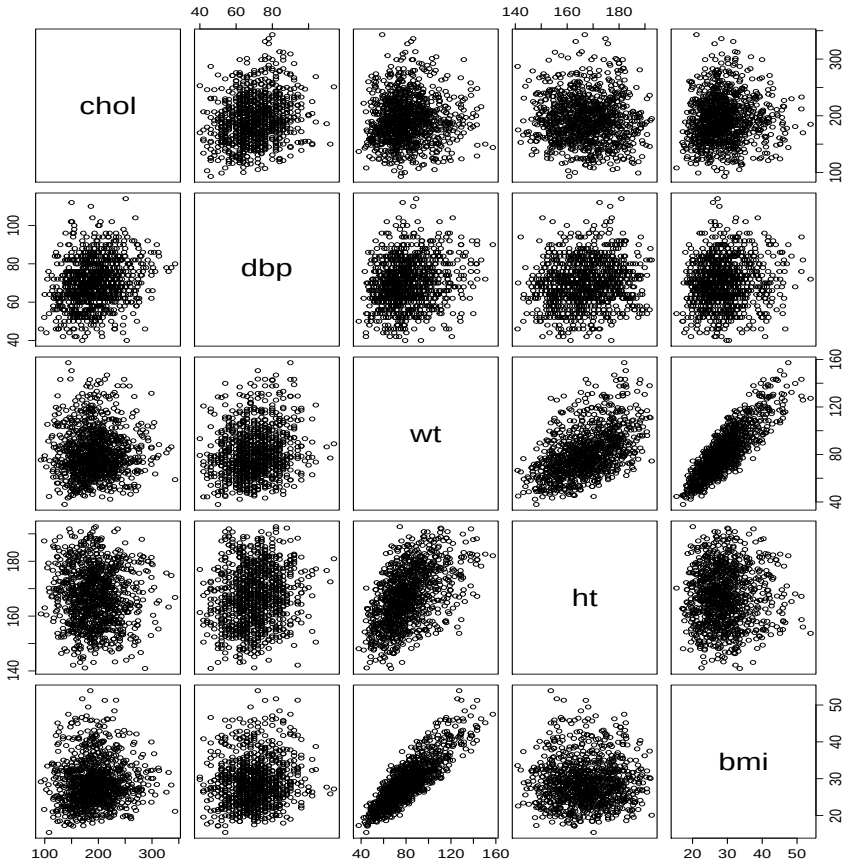

Figure 2: Scatter plots

```
Residuals:
    Min      1Q  Median      3Q     Max
-91.503 -25.689  -2.733  21.014 145.386

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 197.01226   98.64927   1.997   0.0461 *
gender        0.02981    3.61927   0.008   0.9934
age           0.13571    0.08057   1.684   0.0924 .
hdl           0.54652    0.08364   6.534 1.02e-10 ***
sbp           0.02790    0.08128   0.343   0.7315
dbp           0.72565    0.11594   6.259 5.77e-10 ***
wt            0.61421    0.57596   1.066   0.2865
ht           -0.67870    0.58532  -1.160   0.2465
bmi          -1.01345    1.61637  -0.627   0.5308
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 38.4 on 991 degrees of freedom
Multiple R-squared:  0.1069,Adjusted R-squared:  0.09964
F-statistic: 14.82 on 8 and 991 DF,  p-value: < 2.2e-16
```

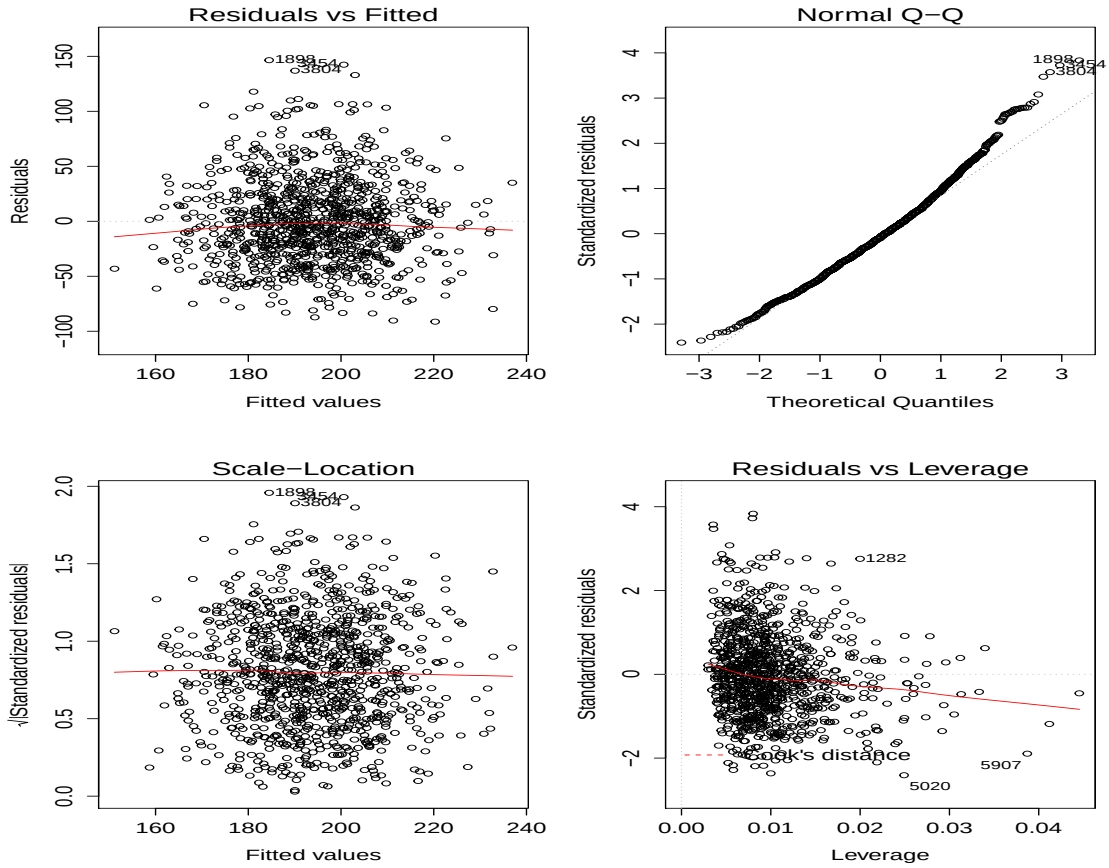

Figure 3: Residual diagnostics

b) Backward model-selection via AIC results in `age + hdl + dbp + wt + ht` as the final model. Comment and discuss.

Solution: hdl and dpb were significant in a) and age was borderline significant. wt and ht were not. However, we know from the scatterplots that wt and ht are strongly correlated AND wt is strongly correlated with bmi. With all these in the model at the same time, collinearity could lead to the significance results of these variables being misleading. That being said, AIC model selection is also affected by collinearity and backward selection is a greedy search. This selected model is somewhat implausible. Is height a predictor of cholesterol level? or could it be that bmi (which is correlated with weight which is correlated with height) is the more logical predictor?

c) We randomly split the data in 50% for training and 50% for testing. We repeat this 100 times, each time recording the model size, the prediction error and the selected model. Below are the results. Comment and discuss.
Average PE Cp: 1505.298, PE AIC: 1505.298, PE BIC: 1510.798
Average Modsize Cp: 4.92, Modsize AIC: 4.92, Modsize BIC: 3.72

```
             modselcp modselaic modselbic
[1,] "gender" "8"      "8"       "0"
[2,] "age"    "58"     "58"      "9"
[3,] "hdl"    "100"    "100"     "100"
[4,] "sbp"    "14"     "14"      "2"
[5,] "dbp"    "100"    "100"     "99"
[6,] "wt"     "21"     "21"      "5"
[7,] "ht"     "13"     "13"      "1"
[8,] "bmi"    "78"     "78"      "56"
```

Solution: AIC and Cp are better in terms of prediction error and pick bigger models than BIC. BIC identifies hdl, dbp and maybe bmi as the predictors of cholesterol (makes sense). AIC and Cp picks these variables and sometimes age as well. wt and ht are rather seldom in the model.

d) I repeat the above using only 25% for training and get the following selection results. Comment and discuss.
Average Modsize Cp: 4.44, Modsize AIC: 4.46, Modsize BIC: 3.08
Average PE Cp: 1525.443, PE AIC: 1525.449, PE BIC: 1536.828

```
             modselcp modselaic modselbic
[1,] "gender" "11"     "11"      "3"
[2,] "age"    "32"     "33"      "5"
[3,] "hdl"    "96"     "96"      "82"
[4,] "sbp"    "12"     "12"      "2"
[5,] "dbp"    "99"     "99"      "93"
[6,] "wt"     "31"     "32"      "5"
[7,] "ht"     "16"     "17"      "4"
[8,] "bmi"    "47"     "46"      "14"
```

Solution: With a smaller training size model selection becomes more unstable. Prediction errors are higher and the models smaller as one would expect the impact of a smaller training data to be. hdl and sbp still stand out as frequently selected variables but now it's not so clear which of the remaining variables to include (if any). Notice also that for AIC and Cp, the wt/bmi choice is not as clear.

e) What would an interaction between gender and age mean? (Come up with a concrete example what such a model might look like and what the coefficients would tell you about subjects in the NHANES data).

Solution: An interaction between gender and age: Let's say we include this in the model and find that the slope of cholesterol on age is larger for men than women. This would tell you that while age leads to increasing cholesterol in general, this increase is even more pronounced among men.

f) A variable "marital status" is also available for analysis. It has 5 levels (single, married, widowed, divorced, separated). Discuss how you would include such a variable in your model. Comment on what kind of results such a model might produce.

Solution: You can code this as 4 dummary variables, picking e.g, single (or the largest group) as baseline. If any of these variables are significant (or selected) it might indicate that marital status is related to cholesterol. Perhaps single people have better(or worse) eating habits etc. Note, this variable is probably related to age as well.

g) Are there any interactions that you would be particularly interesting in estimating? (including with marital status). Discuss and motivate.

Solution: Interactions with marital status. Testing age*marital status could be interesting. Perhaps marital status explains cholesterol differently depending on age. Marital status and gender is also interesting; perhaps being married and being a women and married and being

a man is associated with different levels in cholesterol than say for single men and women? Other interactions that could be interesting would be between the numerical ones. How about hdl and dbp? They are related to cholesterol but is it even worse if both are high? Same for hdl and bmi or dbp and bmi. Also age and bmi - is having a high bmi when older even worse than high bmi when you're young?

Question 2(15=5+5+5)

a) Below you see 4 different nonlinear model descriptions. In each case, identify which parameters are linear and nonlinear. Also identify if there is a data transform that can bring this model into a linear form.
i) $y = a + bx + cx^2 + dx^3$ (parameters a,b,c,d)
ii) $y = a + c * exp(-exp(-b(x - d)))$ parameters(a,b,c,d)
iii) $y = a + b * exp(-cx) + dx$ parameters(a,b,c,d)

Solution: i) is all linear. ii) a and c are linear. iii) a and b and d are linear. For ii, what if you could first estimate the mean of y so that there is no intercept? and scale the data so the c is 1, then log-log would linearize the model in ii.

b) A nonlinear model $y = a + \frac{c}{1+exp(-b(x-d))}$ is fit to a data set of sample size 44 using nonlinear least squares. The regression mean SS is 15.06 and the MSE is 0.13. The F-statistics is thus 114.23. Is this significant? (Say how you conclude this). What does this mean?

Solution: You compare the F-statistics to $F_{4,40}$ and this is highly significant! It means that the model reduced the variance of y to a residual variance (MSE) that is larger than one expects just by chance.

c) The parameter estimates are as follows (as outputted from R):

parameter	estimate	s.e.
b	0.472	0.116
d	27.415	0.551
c	2.617	0.242
d	8.839	0.175

Explain how you would go about setting up CI and assessing significance. What are the assumption you make in order to set up such CI (or compute these p-values)? How would you go about checking those assumptions?

Solution: First, this is a nonlinear fit and so we don't know if the SEs are any good. Check this by looking at the profile plots. If this looks OK you can use an approximate t-based CI. However, you can also use the profile intervals or bootstrap - and you would need to do this if the profile plots indicate that the last linear approximation is not a good approximation near the estimates.

Question 3(20=5+5+5+5)

Data are collected on average race times for 35 different races. Information about the races include the total distance and the overall climb (increase in elevation) during the race.

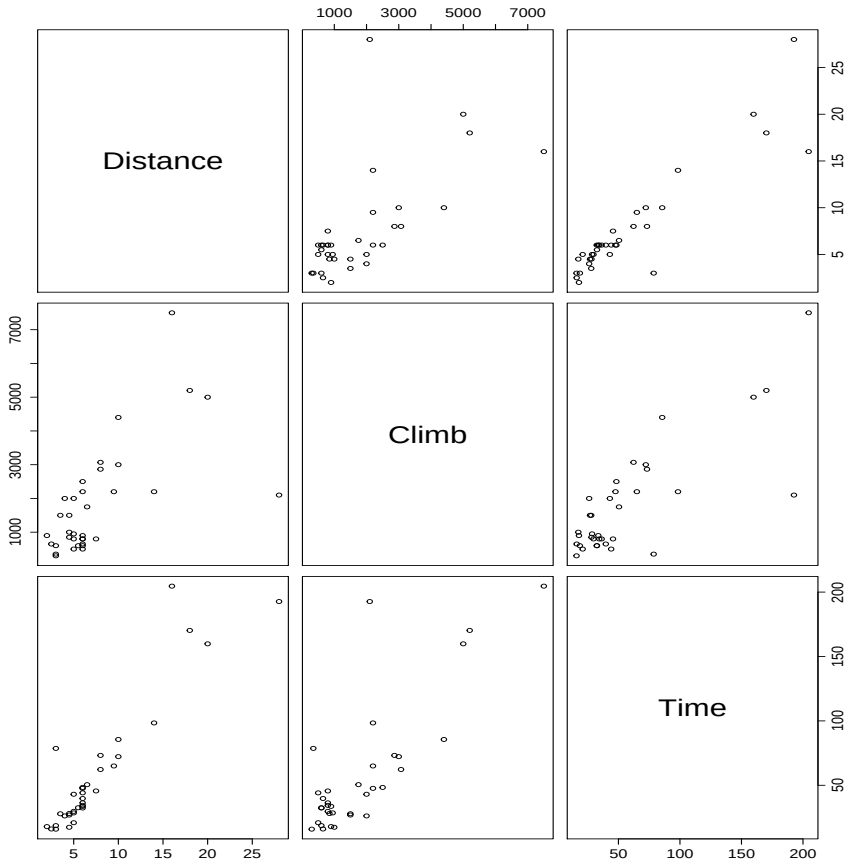

Figure 4: Scatter plots

a) A linear model is fit to the data to predict race times from distance and climb. The results are provided below. Comment on the fit and interpret the model. Any need for transformations? Any outliers? What's the impact on the fit of these/this if any outliers are present? Explain.

Solution: The scatter plots indicate that nonconstant error variance could be a problem, also seen in the diagnostic plot. Perhaps we should log or sqrt the data to even out the spread? We have a couple of outliers (18,7 and 11). One of these (18) has limited leverage but is rather large so will impact the MSE. 11 has high leverage but seems to fit the general trend of the model - might lead to an overly confident R-squared or may have driven the regression fit toward it. 7 is an outlier with high leverage and large residual - this is an outlier that really can have an impact. From the scatter plot distance vs climb we see two extreme leverage observations: one with large distance and climb and one with extreme distance.

The outliers pull the regression slopes upward. I would want to refit without the outliers to see what happens to the fit.

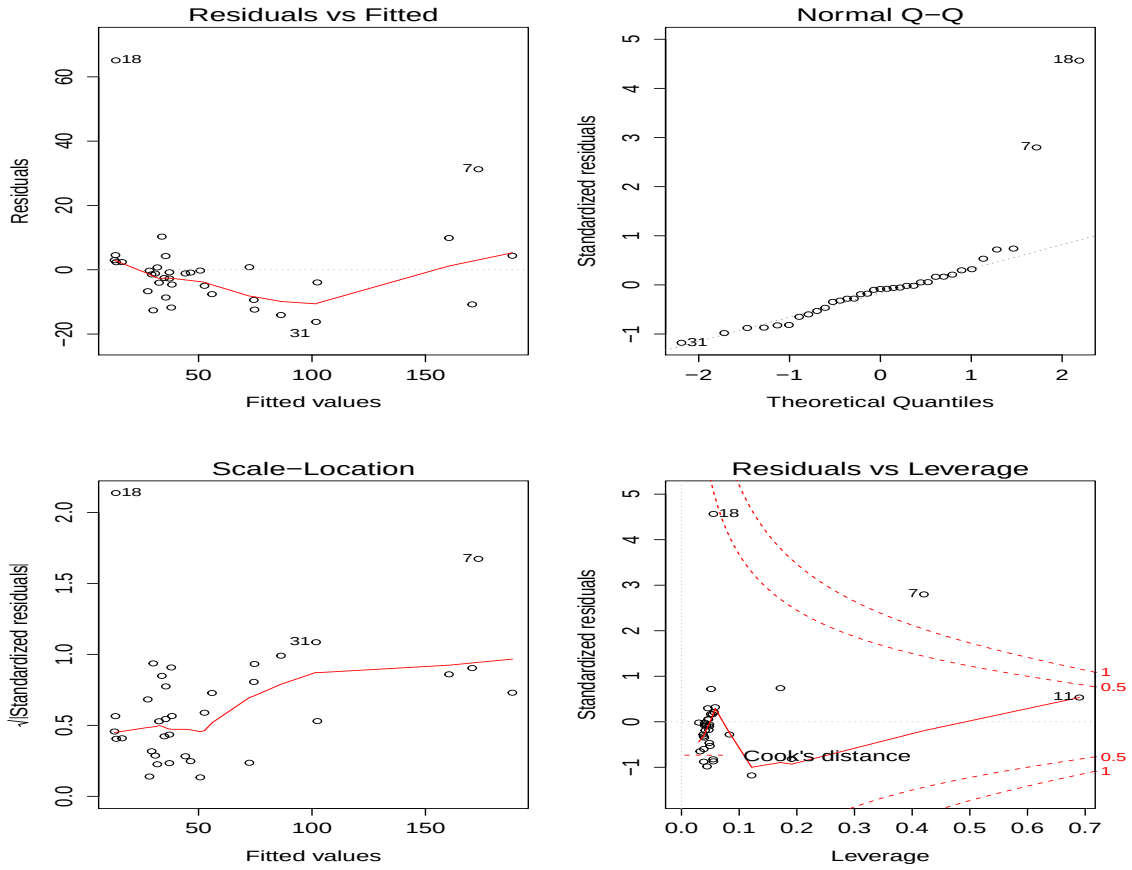

Figure 5: Residual diagnostics

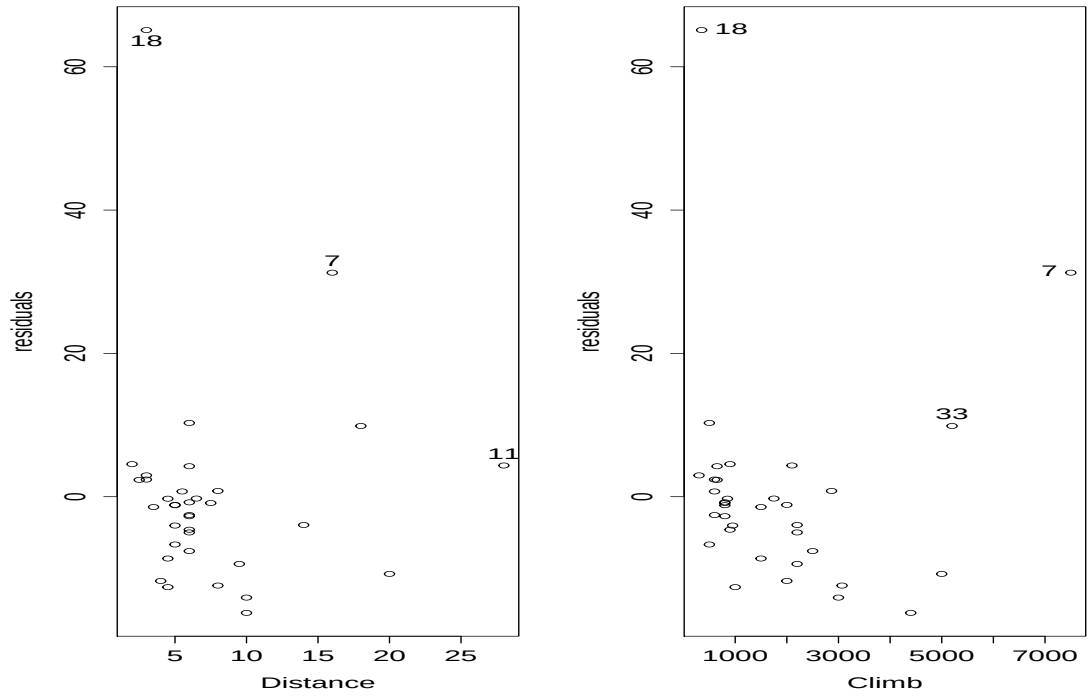

Figure 6: Residuals vs covariates

```
Residuals:
    Min      1Q  Median      3Q     Max
-16.215  -7.129  -1.186   2.371  65.121

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.992039   4.302734  -2.090   0.0447 *
Distance     6.217956   0.601148  10.343 9.86e-12 ***
Climb        0.011048   0.002051   5.387 6.45e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 14.68 on 32 degrees of freedom
Multiple R-squared:  0.9191,Adjusted R-squared:  0.914
F-statistic: 181.7 on 2 and 32 DF,  p-value: < 2.2e-16
```

b) What is the meaning of an interaction term between Climb and Distance? Explain in concrete terms (i.e. not just the statistical definition but what it would mean in this data setting)

Solution: A positive interaction (most likely) between climb and distance would mean that; yes, longer runs take longer and yes, more hilly runs take longer but runs that are both long and hilly take even longer than just the climb and distance suggest separately. Perhaps endurance to the combination of long and hilly races is weaker.

c) Below is the result when an interaction term is included. Comment on the fit. Climb is no longer significant. Comment, interpret and explain.

```
Residuals:
    Min      1Q  Median      3Q     Max
-25.994  -4.968  -2.220   2.381  56.115

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     9.3954374  6.8790233   1.366  0.18183
Distance        4.1489201  0.8352489   4.967 2.36e-05 ***
Climb          -0.0009710  0.0041648  -0.233  0.81718
Distance:Climb  0.0009831  0.0003070   3.203  0.00314 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 12.92 on 31 degrees of freedom
Multiple R-squared:  0.9392,Adjusted R-squared:  0.9333
F-statistic: 159.6 on 3 and 31 DF,  p-value: < 2.2e-16
```

Solution: We didn't center the data so we created a collinearity problem here. This might be why the main effect is no longer significant. (In fact, when you center this data, the interaction is significant and positive).

d) Let's say more data are collected for races at high elevation locations. These appear as a group in the data with longer race times overall and especially for races with larger Climbs. How would you go about modeling this kind of data? (Hint: Draw such a data set in the scatter plots and go from there.)

Solution: At high elevations, it is even more taking to run. You would need to include elevation in the model, perhaps as a numerical variable or, as the question hints, as a dummy variable. You would then need to investigate if elevation has an additive effect on time or if you need to include the interaction with this dummy variable and the Climb and distance.

Question 4(15=5+5+5)

Consider a simple regression model, y on x.
a) When would you consider transforming y? Explain what graphical and numerical tools you would use to determine the need for transformation. Same regarding x.
Solution: Tools: scatter plots and residual plots. I would transform y to even out spread in data or the long-tailed residuals or non-constant residuals. I would transform y and/or x to enhance a linear relationship in a scatter plot. Notice that transformations about y is all about residuals and/or relationship with x. Transformations for x is all about relationship with y or to even out spread.
b) One of the additional assumptions we make in modeling is that of normality. We can use the QQ plot to check this. QQ plots of which of i-iv are relevant for modeling assessment here?
i) QQ of y
ii) QQ of x
iii) QQ of residuals
iv) QQ of standardized residuals
Solution: Number iv is how you check the assumption since standardized residuals have the same SD and so can be compared with one distribution plot. It's only the residuals that need look normal, not x or y (y is normal with different mean values for each observation - will not belong to the same normal distribution so can't be compared with one QQ plot).
c) State an example where a variable transformation can help against collinearity.
Solution: Centering can help when you want to include interactions or polynomial terms in your model. PCA can help in the general case.

Question 5(15=5+5+5)

Below is a scatter plot for a data set.
a) Identify outliers in the data set. Say which measure (leverage, Cook's distance,... that

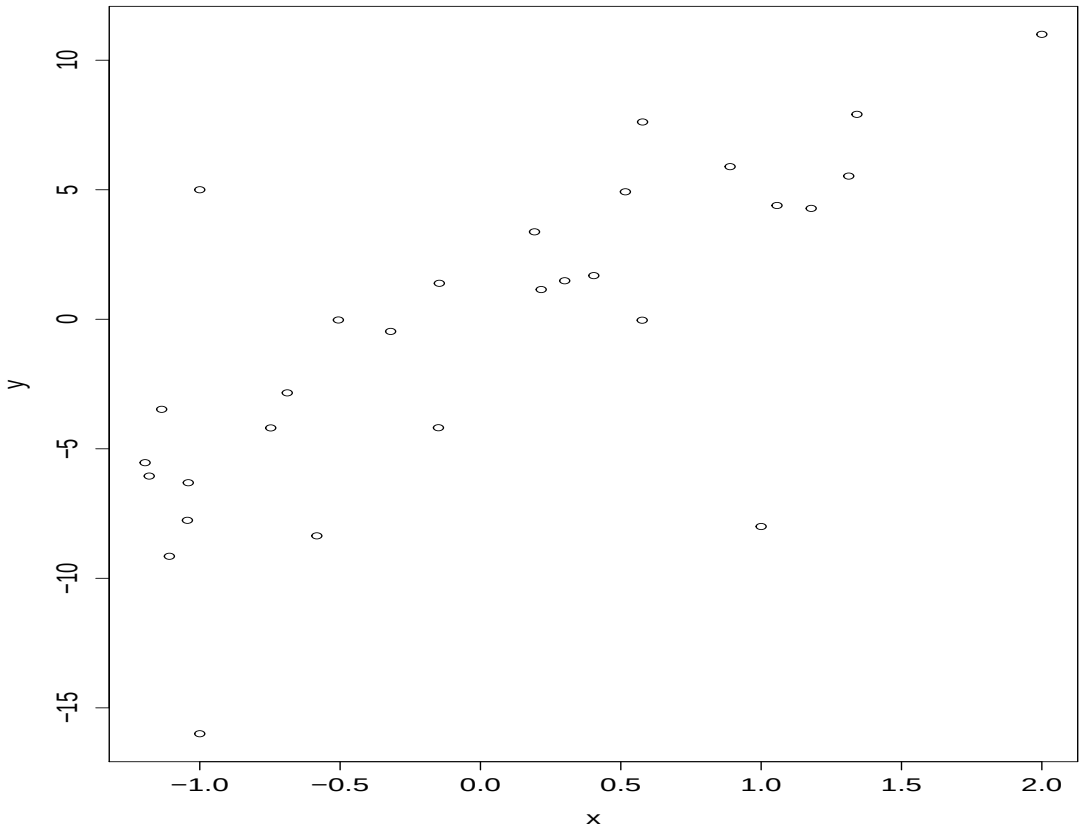

Figure 7: Scatter plot

could be used to identify each of the outliers).
b) Which of R^2, $\hat{\sigma}$, $se(\hat{\beta}_1)$ and the t-value for testing $\beta_1 = 0$ would change if you dropped these outliers (all of them)?
c) Same as b) but you drop only one of the outliers (do this for each of the outliers you identified).

Solution:
a) Outlier 1: (x,y)=(1,-9). Outlier 2: (x,y)=(-1,5). Outlier 3: (x,y)=(-1,-16).
The observation at (2,10) is high leverage.
Outlier 1 would show up in a residual plot and as a high Cook's D. Outliers 2 and 3 have a bit lower leverage than 1 but also large residuals so would also show up in Cook's D and residual plots.
b) If you drop all outliers, the R-squared would increase and the sigma would decrease.

The slope would probably increase a bit without the outliers so the numerator in the t-test would increase, the SE would decrease somewhat (depending on how much sigma decreases compared to the sample size reduction) and so the t would increase (but would also be compared to another t-distribution with 3 fewer degrees of freedom (more long-tailed).
c) Outlier 1: this probably pulled the regression line toward it (down) so the slope would increase and the MSE drop. R-sq and t would increase.
Outlier 2: This has some leverage and if I drop it the whole regression line will probably move downward (intercept) a bit and the slope might increase a bit since outliers 3 and 1 are not balancing each other. MSE will drop a bit.
Outlier 3: If you drop this one, the regression slope may shift even more, both outlier 1 and 2 turn the slope downward.