Violation of Linear Model Assumptions: Non-normal errors, Non-linear conditional mean, unusual observations

Stats 203, Winter 2024

Lecture 11 [Last update: February 26, 2024]

We continue with our discussion of violations of the linear model assumptions. Today we will focus on diagnostics and implications. Next class will discuss remedies.

While there are many different kinds of diagnostics, they mostly center on a common premise: if the linear model with all of its assumptions is correct, then the residuals $e_i = Y_i - \hat{Y}_i$ should look like a collection of (not quite independent) $N(0, \sigma^2)$ random variables. A note: diagnostics are more of an art than a science. The general strategy is to make many plots, and focus only on those which appear suspicious. We will cover only a small subset of the plots one could make: enough to give us a general idea of the most important things to look for, but hardly the full picture.

1 Non-normal errors

Now we suppose all of the assumptions of the linear model hold except Normal errors: that is, $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independently sampled from

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i, \quad \mathbb{E}[\epsilon_i | X_i] = 0, \mathrm{Var}[\epsilon_i | X_i] = \sigma^2,$$

but $\epsilon_i | X_i$ might or might not be Normally distributed.

Implications. In this case our OLS estimator $\hat{\beta}$ will still be unbiased, and the covariance matrix will still be the usual $\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$: this is because the conditional mean and variance of $Y_i | X_i$ are unchanged. But the sampling distribution of $\hat{\beta}$ is not Normal, so our confidence intervals and hypothesis tests – which were based on distributions, like Student's T and the F distribution, derived from the Normal – are not necessarily valid.

Diagnostics: QQ plot. One way to assess whether the errors are Normal is to look at a histogram of the residuals. If the errors are Normal then the residuals are also Normal, and the histogram should look like a bell curve. What we are doing here is visually comparing the PDF of the errors to a Normal PDF. It is also possible to compare the CDFs visually – at first the idea is a little more complicated, but ultimately most people find it more useful.

To diagnose non-Normality in errors, we can sort the residuals in increasing order $e_{(1)} \leq e_{(2)} \leq \ldots \leq e_{(n)}$, and plot the sorted residuals (typically on the y-axis) against the standard Normal quantiles $\Phi^{-1}(\frac{i}{n})$ (typically on the x-axis). This plot is called a quantile-quantile plot, or simply a **QQ plot** for short.

The way to read a QQ plot is this: if the errors are Normal then we expect the relationship between the sorted residuals and Normal quantiles to approximately follow a line. Why? For starters, imagine we had access to the true distribution of ϵ_i, in the form of its CDF F_ϵ. [Of course we don't know the true distribution of ϵ_i, that's what we are trying to figure out, but bear with me.] If $\epsilon_i \sim N(0, \sigma^2)$ were truly Normal, then for any $z \in (-\infty, \infty)$ this CDF would be $F_\epsilon(z) = \Phi((z-0)/\sigma)$, where Φ is the CDF of a standard Normal distribution. You should convince yourself that this implies that the inverse of F_ϵ is simply

$$F_\epsilon^{-1}(\alpha) = \sigma\Phi^{-1}(\alpha), \quad \text{for} \;\; \alpha \in [0,1].$$

Of course $F_\epsilon^{-1}(\alpha)$ is the just the αth quantile of the error distribution. What we have shown is that if the errors are truly Normal then the relationship between $F_\epsilon^{-1}(\alpha)$ and $\Phi^{-1}(\alpha)$ is linear, *regardless of the standard deviation of* ϵ.

Now of course we don't know $F_\epsilon(\alpha)$, because we don't observe the errors, and even if we did observe the errors having n samples does not tell us exactly the distribution. But we do have estimate of F_ϵ: $\widehat{F}_e$, the empirical CDF of the residuals. To be explicit, this is the function given by $\widehat{F}_e(t) = \sum_{i=1}^n \mathbf{1}\{e_i \le t\}$. This function is not invertible – it is just a stepfunction – but for its generalized inverse

$$\widehat{F}^{-1}(p) := \inf\{t : \widehat{F}(t) \le p\}, \tag{1}$$

we have that $\widehat{F}^{-1}(i/n) = e_{(i)}$. Now, assuming $\widehat{F}^{-1} \approx F^{-1}$, then if F is Normal we have

$$e_{(i)} = \widehat{F}^{-1}(i/n) \approx F^{-1}(i/n) \approx \sigma\Phi^{-1}(i/n).$$

So we see that if the errors are truly Normal, then the relationship $e_{(i)}$ and $\Phi^{-1}(i/n)$ should be approximately linear. Of course random fluctuations are expected, but systematic deviations are a sign that something is wrong.

2 Non-linear conditional mean

Now we suppose all of the assumptions of the linear model hold except linearity: that is, $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independently sampled from

$$Y_i = m(X_i) + \epsilon_i, \quad \mathbb{E}[\epsilon_i | X_i] = 0, \mathrm{Var}[\epsilon_i | X_i] = \sigma^2,$$

but $m(x) := \mathbb{E}[Y|X = x]$ might or might not be a linear function of X_i.

Implications. As you might imagine, given that linear is in the name of the model, having a non-linear conditional mean is a **problem**. In particular, it is difficult to talk about the accuracy of the OLS estimator since we don't even know what parameters estimating! It turns out that there is still a way to make sense of OLS even when the conditional mean is not linear, which we will get to when we discuss remedies.

Diagnostics: added variable plots. A natural thing to do would be to plot the vector of responses $(Y_1, \ldots, Y_n)$ against the jth predictor $(\mathbf{X}_{1,j+1}, \ldots, \mathbf{X}_{n,j+1})$ and see whether the relationship appears linear. This is useful and typically we do it before we even run a regression. However these plots assess the *marginal* relationship between the response and each predictor, rather than the relationship after we have "adjusted for" all the other predictors: the latter, of course, is what we are truly after.

Another useful diagnostic is a to plot the fitted values $\hat{y}$ against the residuals e. We have already seen how this is useful in identifying the presence of heteroskedastic errors, but it can also be used to diagnose non-linearity. The problem is that this does not tell us which predictor(s) have a non-linear relationship with Y.

A more informative set of diagnostics is are **added variable plots**. These plots tell us the relationship between Y and predictor j after "adjusting for" the effects of all other predictors. They are created using the following three steps.

1. Let $\mathbf{X}_{-j} \in \mathbb{R}^{n \times p}$ be the matrix with all predictors except the jth predictor, and let $\mathbf{H}_{-j}$ be the corresponding hat matrix. Run ordinary least squares with Y as the response and $\mathbf{X}_{-j}$ as the predictor matrix. Let $e_j = (\mathbf{I} - \mathbf{H}_{-j})Y$ be the residuals in this regression.

2. Let $x_{(j)} \in \mathbb{R}^n$ be the vector of observed values of the jth predictor. Run ordinary least squares with $x_{(j)}$ as the response and $\mathbf{X}_{-j}$ as the predictor matrix. Let $r_j = (\mathbf{I} - \mathbf{H}_{-j})x_{(j)}$ be the residuals in this regression.
3. Plot the residuals of the two regressions against one another, with r_j on the x-axis and e_j on the y-axis. Plot a line through the original with intercept equal to 0 and slope equal to $\hat{\beta}_j$.

Applying the above steps to each predictor $x_{(1)}, \dots, x_{(p)}$ will yield p different scatterplots. If the conditional mean function is truly linear, then we expect each of these plots to look like a cloud around a line. In fact the slope of the line will be $\hat{\beta}_j$, as you showed in Problem 3, Homework 2. If the conditional mean function is *not* linear, the added variable plots can help isolate which variable(s) are driving the non-linearity.

Added variable plots are a very neat idea in that they take advantage of the "partial regression" interpretation of multiple regression. In theory, they can also be useful for detecting violations such as non-constant variance errors, outliers, etc. Nowadays, I think most statisticians would say that they are rarely the most useful diagnostic.

3 Correlated errors

I will not say much about the case where $(X_1, Y_1), \dots, (X_n, Y_n)$ are not independent. One thing worth pointing out is that all that is really needed is for $Y_1, \dots, Y_n$ to be mutually independent given $\mathbf{X}$: as usual, assumptions on the distribution of $\mathbf{X}$ are unimportant, it's the assumptions we make on $Y|\mathbf{X}$ that matter. In general, however, even assessing whether $Y|\mathbf{X}$ are independent is hard to diagnose and even harder to fix, since there are so many ways in which the errors (and thus the responses) can be correlated.

See Faraway 6.1.3 for an example of diagnosing temporal correlations in the errors.

4 Unusual observations and outliers

Finally, we discuss unusual observations and outliers. These are different than the other violations of the linear model. So far we have assumed that all data are observed values of random variables distributed according to a regression model, but that the regression model fails to satisfy one or another of the assumptions of the linear model.

Now we assume that the vast majority of points do follow the linear model, but we are worried that there may be a few aberrant data points. In fact, even *one* unusual observation can be bad enough to throw off the whole OLS operation.

There are several ways a point can be unusual. Namely, it can have an unusual x value, in which case it is called **leverage point**. Or it can have an usual y value, in which case it is called an **outlier**. Or it can have a large **influence** on the regression, in that if we deleted the point then (some aspect of) the regression would changed.

4.1 Outliers

An outlier is an observation with a response value that is very different than what would be expected given its predictors. The residual $e_i = y_i - \hat{y}_i$ seems like a natural measure of the "outlyingness" of an observation. If the squared residual e_i^2 is large then we suspect the point is an outlier.

Obviously, the expected value of e_i^2 depends on σ^2, and we need to account for this when deciding whether or not a given observation has a "suspiciously" large residual. One way of doing this is to instead look at the **standardized residual**:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - \mathbf{H}_{ii}}}.$$

If the linear model is correct and all observations follow the model, then r_i should be approximately standard Normal distributed. Thus, by comparing the standardized residuals to quantiles of the standard Normal distribution, we get a sense of how unlikely they would be if all data in fact came from a linear model.

Jackknife residuals. This is a good idea but it has a slight problem: if (x_i, y_i) is in fact an outlier then we do not want to use it to estimate the parameters $\hat{\beta}$ or $\hat{\sigma}$. A better idea is to use all the data *except* (x_i, y_i) to fit the regression, and then see how y_i compares to the predicted value of this "leave-one-out" regression.

Let $\mathbf{X}_{(i)} \in \mathbb{R}^{(n-1)\times(p+1)}$ be the predictor matrix with the ith observation excluded, and let $\hat{\beta}^{(i)}$ be the resulting OLS estimates. The resulting regression will predict that the response at x_i is $\hat{Y}_{(i)} := \hat{\beta}_0^{(i)} + \sum_{j=1}^{p} \hat{\beta}_j^{(i)} x_{ij} = (1\ x_i^\top)\hat{\beta}^{(i)}$. Note that is not the fitted value $\hat{y}_i$, since observation i was excluded from the regression. The difference between the actual response and the predicted response is $Y_i - \hat{Y}_{(i)}$, and this has variance

$$\begin{aligned}
\mathrm{Var}[Y_i - \hat{Y}_{(i)}|\mathbf{X}] &= \mathrm{Var}[Y_i|\mathbf{X}] + \mathrm{Var}[\hat{Y}_{(i)}|\mathbf{X}] \\
&= \sigma^2 + \mathrm{Var}[(1\ x_i)^\top \hat{\beta}_{(i)}|\mathbf{X}] \\
&= \sigma^2 + \sigma^2 (1\ x_i)^\top (\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1} (1\ x_i).
\end{aligned}$$

It follows from this that if all of the assumptions of the linear model are correct, and all of the data come from the linear model, then

$$t_i := \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)} \sqrt{1 + (1\ x_i)^\top (\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1} (1\ x_i)}} \sim t_{n-1-(p+1)}.$$

In statistics the procedure of computing estimators on leave-one-out datasets is called the **jackknife**. Thus, t_i are called the jackknife residuals. Comparing jackknife residuals to the quantiles of a Student T distribution gives a sense of how unlikely an observed response Y_i would be if the linear model were in fact correct. This is the way we diagnose outliers.

On its face, in order to compute $t_1, \ldots, t_n$ we need to run n-regressions. If n is large, this can be time consuming, even for `R`. Luckily there is a **shortcut formula**:

$$t_i^2 = r_i^2 \left(\frac{n-p-2}{n-p-1-r_i^2} \right)$$

Thus, can compute the jackknife residuals from the standardized residuals the regression with all of the data included.

In every residual plot we have looked at so far, it can also be useful to plot standardized or jackknife residuals.

Correcting for multiple comparisons. What quantile of the T distribution should we compare $|t_1|, \ldots, |t_n|$ to? There is an issue here involving multiple comparisons. Suppose we had $n = 100$ observations, and compared $|t_i|$ to $t_{n-p-2}^{1-.05/2}$, the 97.5th percentile of a Student T distribution with $n-p-2$ degrees of freedom. Even if every observation came from the linear model, we would expect about $2.5 + 2.5 = 5$ observations to have absolute standardized residuals above $t_{n-p-2}^{.975}$.

The issue here is that we are looking at many data points, and so some will happen to appear unusual just by chance. To correct for this, we need to use a different quantile. A common suggestion is to use the quantile $t_{n-p-2}^{1-\alpha/2n}$. So, in the previous example, we would use $t_{n-p-2}^{1-.025/100} = t_{n-p-2}^{.99975}$. To see why this works, suppose the linear model is correct and there are no true outliers. Let O_i be the event that we (falsely) flag observation i as an outlier, that is, that $|t_i| > t_{n-p-2}^{.99975}$. The probability that we falsely identify any point as an outlier is at most

$$\begin{aligned}
\mathbb{P}\Big(\bigcup_{i=1}^{n} O_i \Big) &\leq \sum_{i=1}^{n} \mathbb{P}(O_i) \\
&\leq \sum_{i=1}^{n} \alpha/n, \\
&= \alpha.
\end{aligned}$$

This approach to correcting for multiple comparisons is known as the **Bonferonni correction**, and it applies in many situations besides outlier detection.

4.2 Leverage points

Leverage points are observations that have the *potential* to highly influence the OLS estimators, based on the values of their predictors. This is distinct from observations that actually *do* influence the OLS estimators, which are covered in the next section.

Mathematically the leverage of observations i is the ith diagonal element of the hat matrix: recalling that the ith row of $\mathbf{X}$ is $\mathbf{X}_{i\cdot} = (1\ x_i)^\top$

$$\mathbf{H}_{ii} = [\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]_{ii} = (1\ x_i)^\top(\mathbf{X}^\top\mathbf{X})^{-1}(1\ x_i).$$

Notice that this definition does not depend on Y.

To motivate this definition, recall that $\mathrm{Var}(e_i|\mathbf{X}) = \sigma^2(1 - \mathbf{H}_{ii})$. This implies that the leverage is between 0 and 1. The closer the leverage is to 1, the smaller the expected squared residual e_i^2, and the more $\hat{y}_i$ is "forced" to be near y_i.

Which points have large leverage? Using block matrix inversion, it can be shown that

$$\mathbf{H}_{ii} = \frac{1}{n} + (x_i - \bar{x})^\top \widehat{\mathbf{\Sigma}}^{-1}(x_i - \bar{x}),$$

where $\widehat{\mathbf{\Sigma}}$ is the empirical covariance of the predictors, and $\bar{x} \in \mathbb{R}^p$ is the empirical means of the predictors. The second term above is called the **Mahalanobis distance**, in this case between x_i and $\bar{x}$. The interpretation is that the further x_i is from the empirical average of the predictors, the more leverage it holds over the regression.

Notice that nothing in the assumptions of the linear model rules out leverage points, since they depend only on the predictors. Why, then, do we care about large leverage? Roughly speaking, the existence of a high-leverage point X_i means that the linear model can potentially be highly influenced by the corresponding Y_i. Thus the OLS estimates $\hat{\beta}$ and fitted values $\hat{Y}$ may be highly unstable (meaning they can have high variance.) As a result the standard errors of $\hat{\beta}$ may be large, and the Wald intervals may be very wide.

4.3 Influence points

More concerning than high leverage is if a given observation *actually* changes the regression. We can measure this by comparing the OLS coefficients $\hat{\beta}$ to the leave-one-out coefficients $\hat{\beta}_{(-i)}$. One way to summarize the difference between these two vectors is **Cook's distance**. Cook's distance compares the fitted values $\hat{Y}$ to the predictions when observation (X_i, Y_i) is left out:

$$D_j = \frac{1}{(p+1)\hat{\sigma}^2}\sum_{i=1}^{n}(\hat{Y}_i - \hat{Y}_{(j)i})^2.$$

The notation $\hat{Y}_{(i)j} = (1\ x_j)^\top \hat{\beta}^{(i)}$ means take the ith fitted value from the regression with the jth observation dropped. The Cook's distance satisfies the relation

$$D_i = \Big(\frac{r_i^2}{p+1}\Big)\Big(\frac{\mathbf{H}_{ii}}{1 - \mathbf{H}_{ii}}\Big)$$

Remember that r_i^2 measures whether a point is an outlier, and $\mathbf{H}_{ii}$ whether a point has leverage. Thus, we see that the Cook's distance measure of influence is measuring a combination of outlyingness and leverage.

Large values of D_i indicate the regression changes quite a bit when (X_i, Y_i) are omitted. A rule of thumb is that a given Cook's distance is notable if $D_i > 1$.

Diagnostics with Scottish hill races

We go over diagnostics in R. The data are $n = 35$ record times for different Scottish hill races.

```r
url = 'http://www.statsci.org/data/general/hills.txt'
races = read.table(url, header=TRUE, sep='\t')
head(races)
```

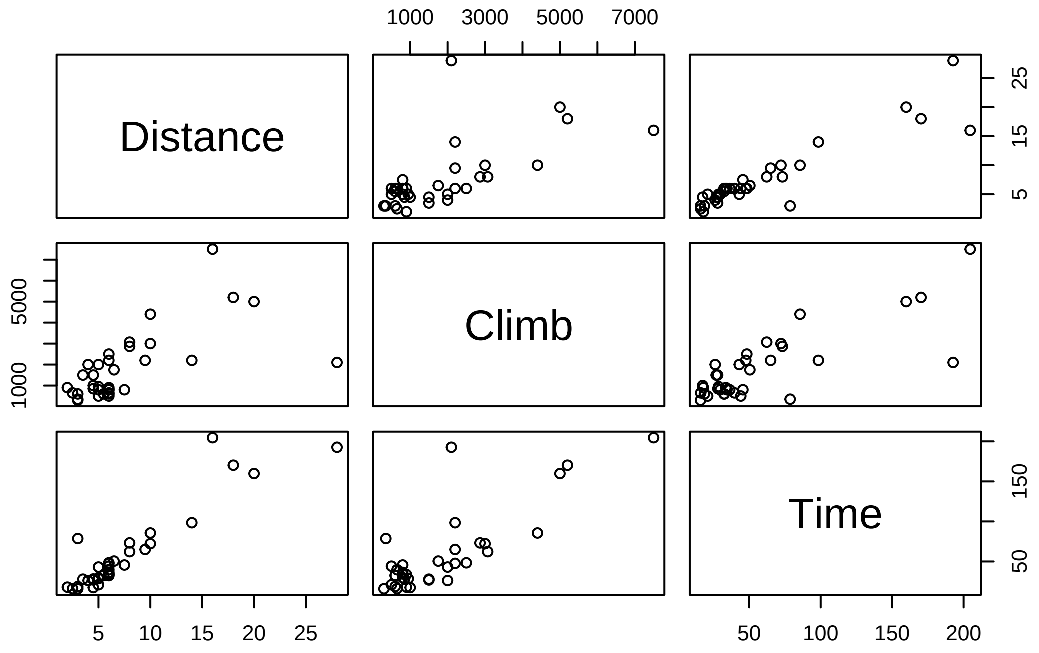
```
##          Race Distance Climb   Time
## 1 Greenmantle      2.5   650 16.083
## 2    Carnethy      6.0  2500 48.350
## 3 CraigDunain      6.0   900 33.650
## 4      BenRha      7.5   800 45.600
## 5   BenLomond      8.0  3070 62.267
## 6    Goatfell      8.0  2866 73.217
```

For exploratory data analysis, we plot and summarize the data, omitting the first column which is just the name of the race.

```r
summary(races[,-1])
```

```
##     Distance          Climb           Time       
##  Min.   : 2.000   Min.   : 300   Min.   : 15.95  
##  1st Qu.: 4.500   1st Qu.: 725   1st Qu.: 28.00  
##  Median : 6.000   Median :1000   Median : 39.75  
##  Mean   : 7.529   Mean   :1815   Mean   : 57.88  
##  3rd Qu.: 8.000   3rd Qu.:2200   3rd Qu.: 68.62  
##  Max.   :28.000   Max.   :7500   Max.   :204.62  
```

```r
plot(races[,-1])
```



Linearity is plausible, although there are clearly a few observations with unusual predictors, responses, or both. So we fit a model to predict record time (in minutes) from distance covered (in miles) and elevation gained (in feet).

```
races.lm = lm(Time ~ Distance + Climb, data = races)
summary(races.lm)
```

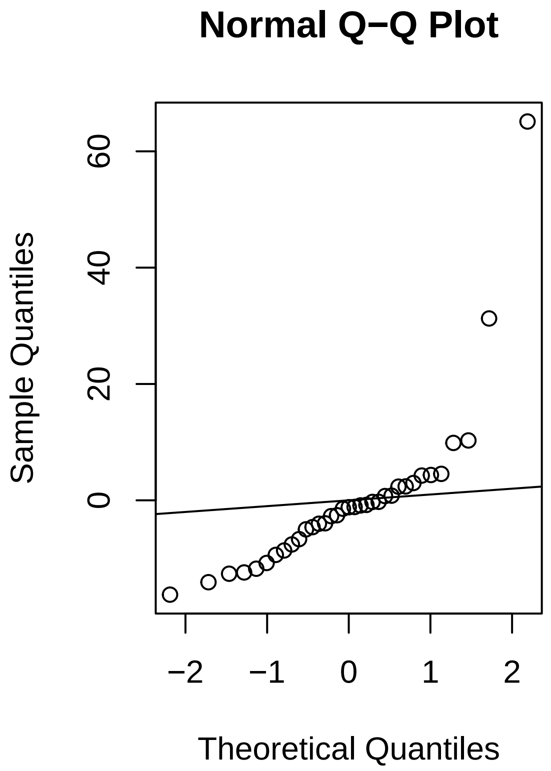
```
##
## Call:
## lm(formula = Time ~ Distance + Climb, data = races)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.215  -7.129  -1.186   2.371  65.121
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.992039   4.302734  -2.090   0.0447 *
## 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
```

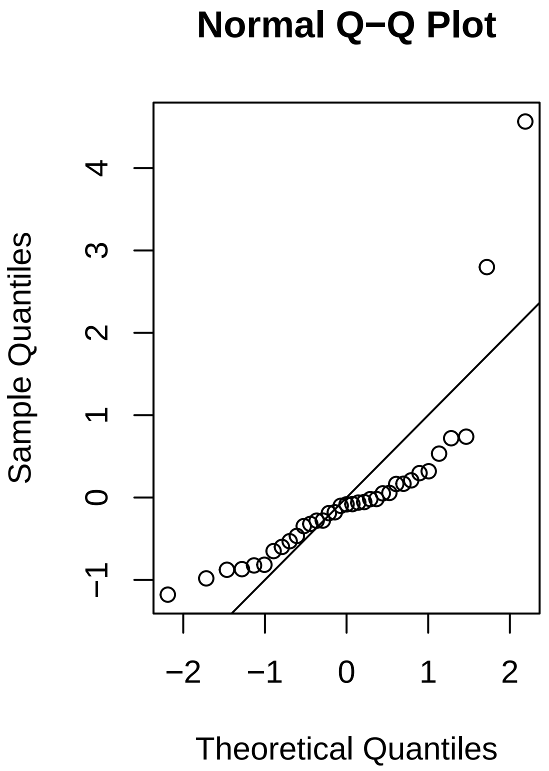
Both predictors seem positively associated with record time (conditional on the other), which makes perfect intuitive sense. The association is highly statistically significant, *if the linear model is correct.* But is the model correct?

QQ plot

To check whether the error distribution is truly Normal, we compare the observed residuals to the quantiles of a Normal distribution, using a QQ-plot. We can plot any of the raw residuals, the standardized residuals, or the jackknife residuals. In this case we plot the first two.

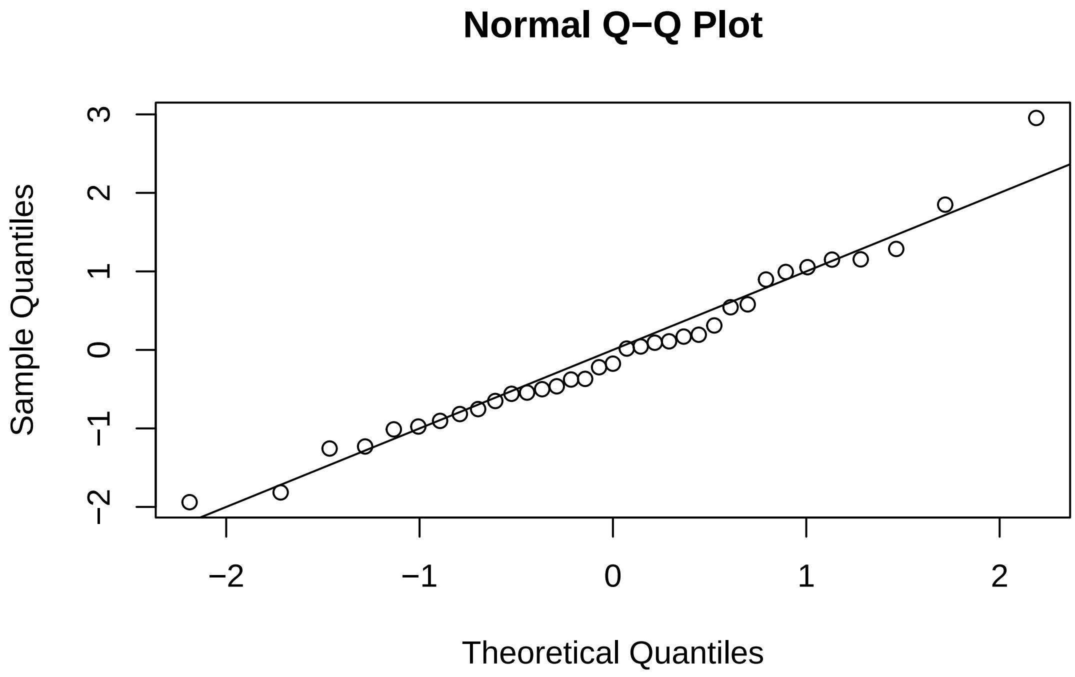
```
par(mfrow = c(1,2))
qqnorm(resid(races.lm)) # raw residuals
abline(a = 0,b = 1)

qqnorm(rstandard(races.lm)) # standardized residuals
abline(a = 0,b = 1)
```





The distribution of residuals seems far from Normal. We know this since if the residuals were truly Normal we'd expect the points to roughly follow the diagonal, as in this simulated example:

```
simulated_errors = rnorm(35)
qqnorm(simulated_errors)
abline(0, 1)
```



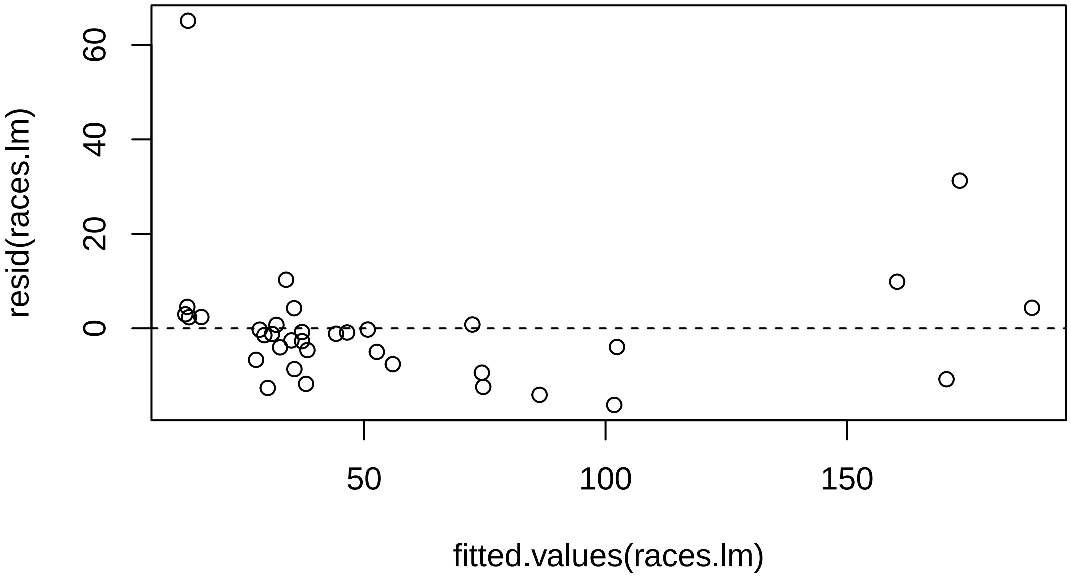
Note that even in the perfectly Normal case, there is some deviation from a perfect linear relationship due to random chance, particularly for very small and very large quantiles. But the deviation is quite a bit less than we see for the Scottish hill races.

Heteroskedasticity

Next we check whether the errors have constant variance. We can plot any or all of the raw, standardized, or jackknife residuals, or absolute residuals, against the fitted values.

Here we just look at the raw residuals.

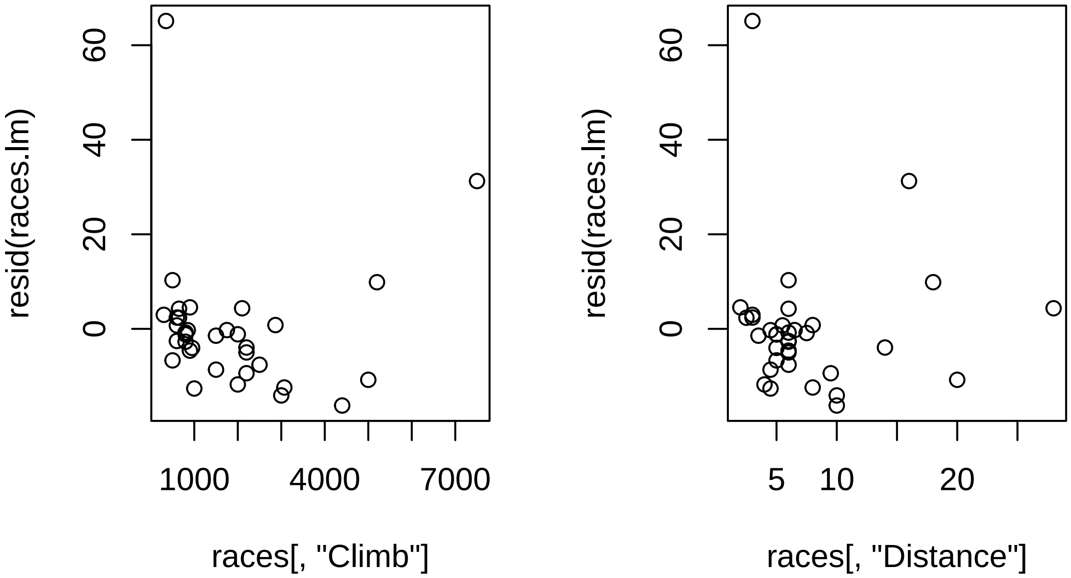
```
plot(fitted.values(races.lm),resid(races.lm))
abline(h=0, lty=2)
```



We see that the spread in the residuals increases with the fitted values. This makes sense since the response must be positive. We also see a clear trend in the conditional mean of the residuals, indicating that the conditional mean $m(x)$ is possibly non-linear. Finally, we see one potential outlier in the top-left corner.

It is similarly valuable to plot the residuals against each of the predictors individually. In this case, the plots tell a similar story.

```r
par(mfrow = c(1,2))
plot(races[,"Climb"],resid(races.lm))
plot(races[,"Distance"],resid(races.lm))
```



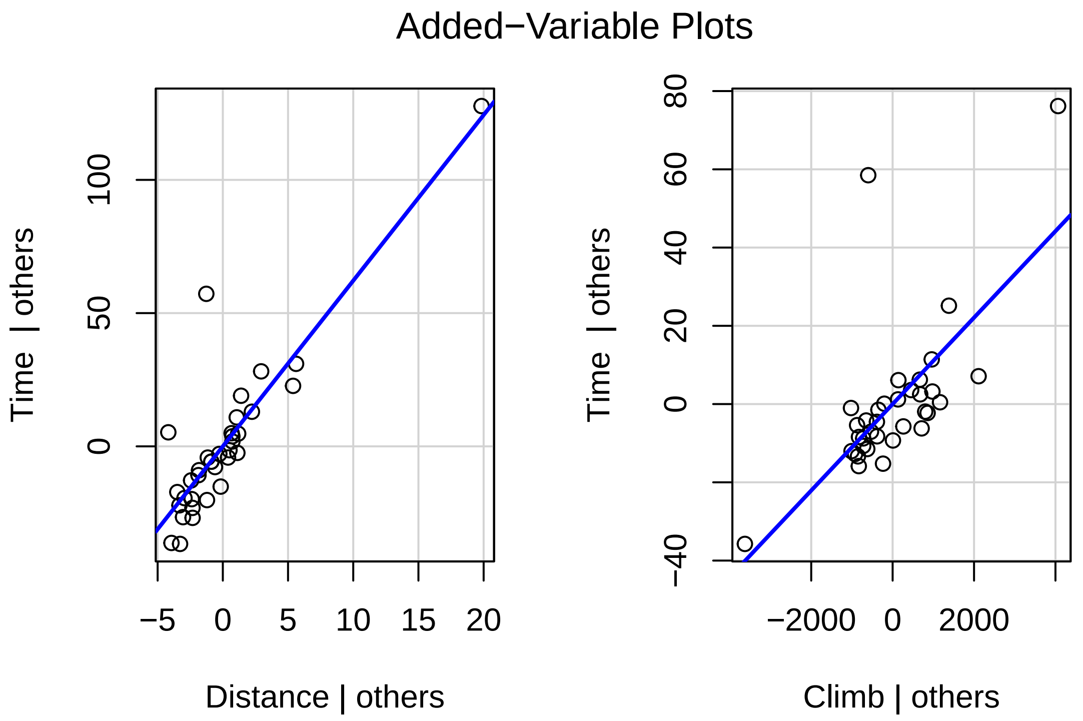
Non-linearity

The `car` package in `R` has a function `avPlots` that directly computes added variable plots.

```r
library(car)
```

```
## Loading required package: carData
```

```r
avPlots(races.lm, id = F)
```



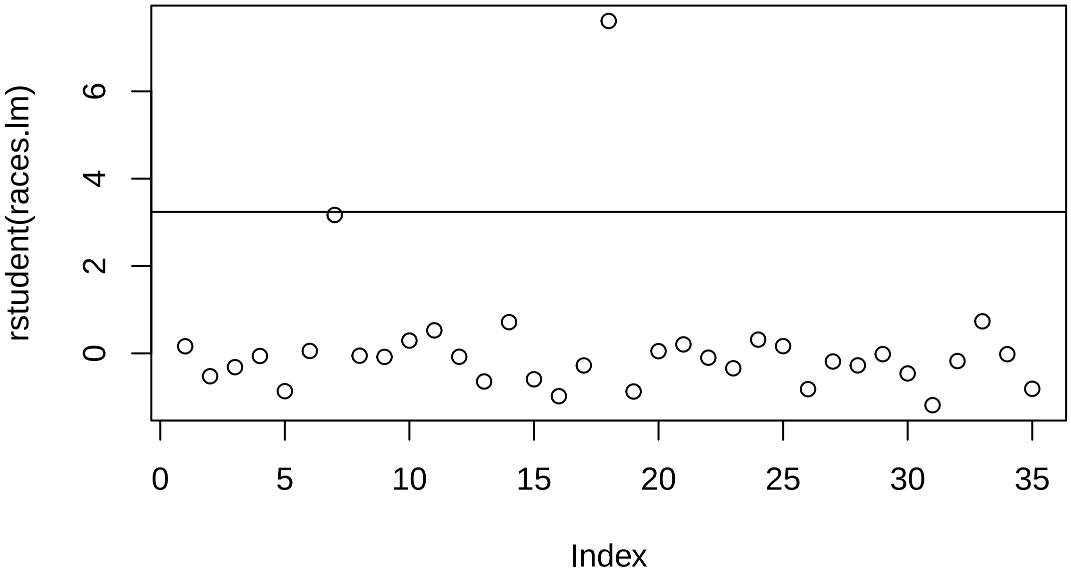
The "|others' ' labeling indicates that the x and y-axes represent the residuals of the variable in question, after partialing out the other predictors. The `id = F` flag prevents the plots from identifying the points with the largest leverage and residuals.

The added variable plots suggest that the conditional mean appears to be reasonably close to linear, but with a few fairly obvious outliers. We turn to this next.

Outliers

The `rstudent` function computes jackknife residuals. We could use jackknife residuals in any of our previous plots as a way to check for violations. If we are purely looking for outliers there is no obvious other variable to plot against. We can, if we want, just plot them in lexical order.

```
p = 2           # Number of predictors
n = nrow(races) # Number of observations
alpha = 0.1     # Significance level
cutoff = qt(1 - alpha/(2*n),df = n - p - 2) # = Bonferroni corrected quantile

plot(rstudent(races.lm))
abline(h = cutoff) # Horizontal line at appropriate T quantiles
abline(h = -cutoff) # Horizontal line at appropriate T quantiles
```



```
races[abs(rstudent(races.lm)) > cutoff,]
```

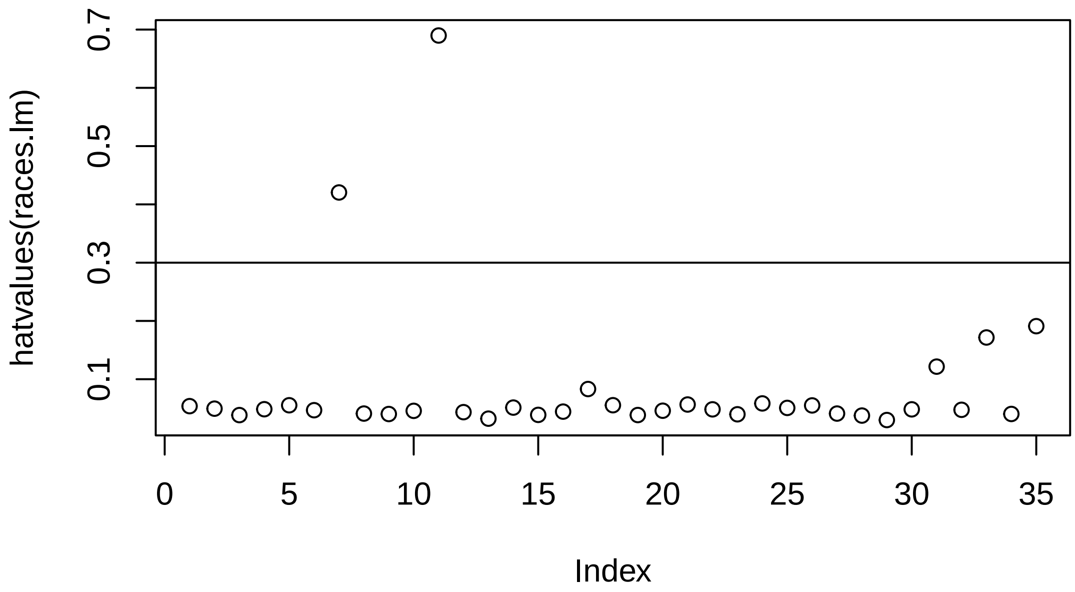
```
##          Race Distance Climb  Time
## 18 KnockHill        3   350 78.65
```

One observation has been identified as an outlier: the Knock Hill race. It turns out that this is a known error.

Leverage

The `hatvalues` function in `R` computes leverages $\mathbf{H}_{ii}$. There is no obvious quantity to plot leverage against. Again we just plot them in lexical order.

```
plot(hatvalues(races.lm))
cutoff = .3
abline(h = cutoff)
```



```
races[hatvalues(races.lm) > cutoff,]
```

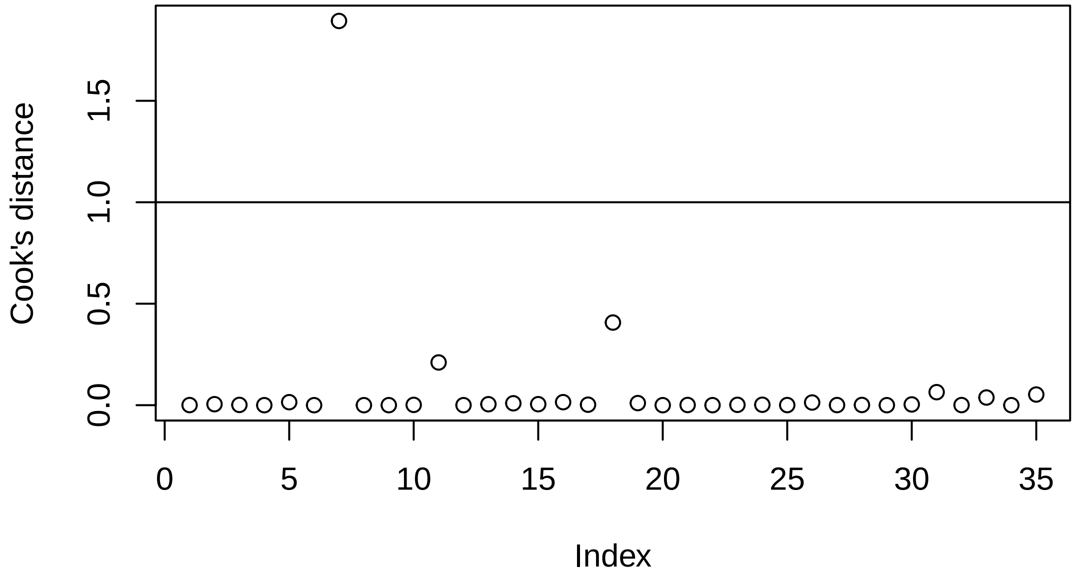
```
##           Race Distance Climb    Time
## 7  BensofJura       16  7500 204.617
## 11 LairigGhru       28  2100 192.667
```

We see that there are two high-leverage points. Both have unusual predictors: `BensofJura` has a lot of elevation gain, while `LairigGhru` is a very long race.

Influence

Finally, to assess influence, we plot Cook's distance, again in lexical order.

```
cutoff = 1
plot(cooks.distance(races.lm),ylab = "Cook's distance")
abline(h = cutoff)
```



```
races[cooks.distance(races.lm) > cutoff,]
```

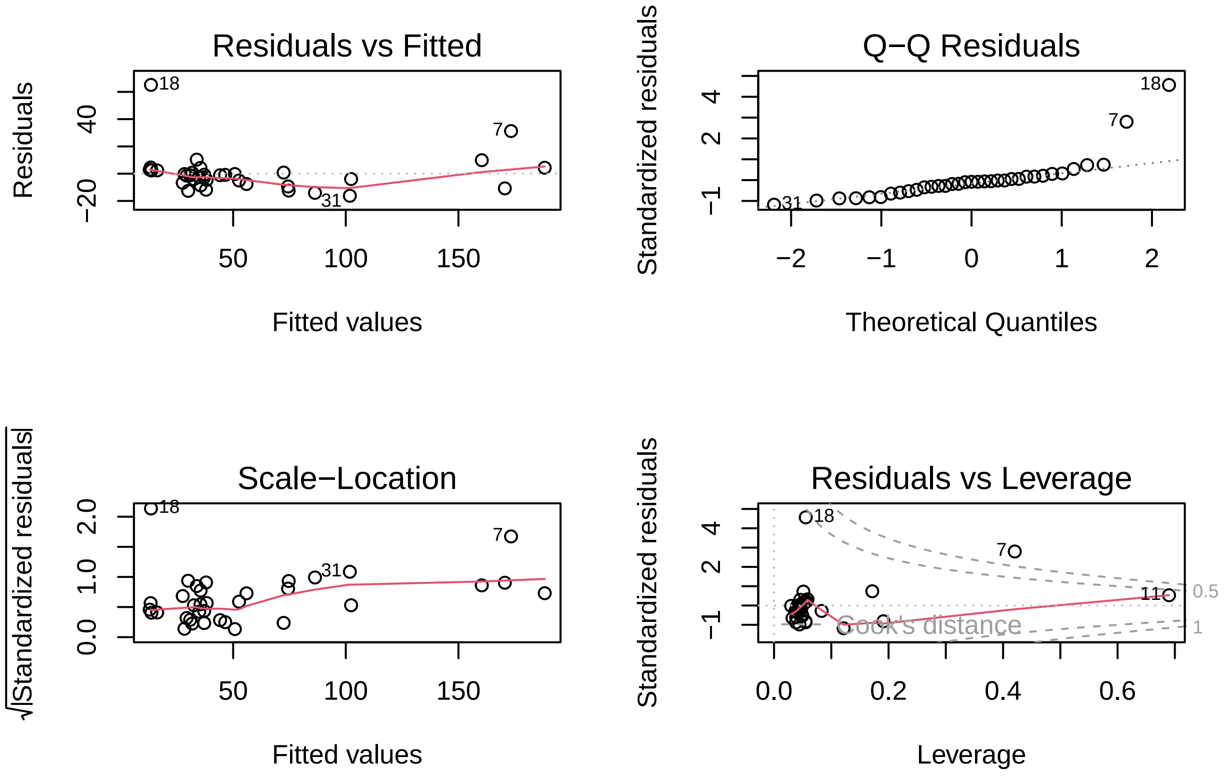
```
##         Race Distance Climb    Time
## 7 BensofJura       16  7500 204.617
```

Of our three unusual observations – one outlier, two leverage points – only `BensofJura` has a Cook's distance larger than 1.

The default plots

Finally, it is possible to directly call `plot` on an `lm` object.

```
par(mfrow = c(2,2))
plot(races.lm)
```



This produces 4 plots. The first is our usual plot of the raw residuals against the fitted values. The second is a QQ plot using standardized residuals. The third plots absolute standardized residuals against fitted values. We've seen each of these before.

The last is new. It plots leverage on the x axis and standardized residuals on the y axis. The dashed curves show level sets of Cook's distance – that means any two points on the same dashed curve have the same Cook's distance. We have seen that influence, in the form of Cook's distance, will be large for points that have high leverage and large absolute (jackknife) residuals. Thus, points towards the top right or bottom right of the plot will be high influence points.

In each of these plots points with high leverage, standardized residuals, or both are explicitly marked. `R` has its own rules for determining cutoffs that are similar to the ones we've gone over in class.