Stat 414 - Day 2
Basic Regression Model Assumptions

Last Time:

A regression equation of the form $y = \hat{\beta}_0 + \hat{\beta}_1 x$ has intercept $\hat{\beta}_0$ and slope coefficient $\hat{\beta}_1$. We can think of the model as connecting the expected values of the populations of responses at each x value. We assume these populations are normally distributed and all have the same variability σ^2.

Let y_i represent the speed of the winning horse in year i.
Model 1: $Y_i = \beta_0 + \beta_1 Year_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$.

Additional assumptions include:

- Independence in the errors (e.g., no time dependence, no clustering)
- X values are ``fixed'' and measured without error

Residual plots are one way to check some of these assumptions. We want

- Residuals vs. fitted values to show no curvature (satisfying linearity)
- Residuals (after subtracting off the means) to follow a (roughly) normal distribution
- Residuals vs. fitted values to show no fanning or heterogeneity

When these assumptions are not met, common remedies are transformations and polynomial models. See also splines and generalized additive models.

Example 1: Kentucky Derby cont.

We previously saw that the relationship between speed and year was not linear.

Quadratic model

A quadratic model includes both $year$ and $year^2$ in the model. The additional term allows the model to "turn."

```
#Create the quadratic term
yearsq = KYDerby23$Year*KYDerby23$Year

#We can also use the I() function to tell R to evaluate the expression before
fitting the model
model2 = lm(speed~Year + I(Year^2), data = KYDerby23)
model2
```

(a) The coefficient of $year$ is positive and the coefficient of $year^2$ is negative. What does this imply about the behavior of the model?

Does the model seem to have the right form?

```
plot(KYDerby23$speed~KYDerby23$Year)
lines(cbind(KYDerby23$Year, model2$fitted.values), col="red")
```

Are the model conditions more adequately met?

(b) Which of the regression model conditions/plots changed? Improvement?

Log transformation

If we consider the relationship "monotonic" with Y increasing at a slower and slower rate, we can try a log transformation of the X variable (to "slow it down").

```
#Like many other packages "log" refers to natural log
log.year = log(KYDerby23$Year)
model3 = lm(speed ~ log.year, data = KYDerby23)
```

This model does not appear to be very helpful! The model we are fitting is curved, but not curved in the right place. We can often solve this by first shifting the data…

```
#Let's make the first year = 1 (we could start at zero but then couldn't take the
log)
shiftedyear = KYDerby23$Year - 1874
logx = log(shiftedyear)
model3b = lm(speed~logx)
plot(speed~logx)
```

(c) Is the association between speed and log(year) linear?

Does the model seem to have the right form?

(d) Are the model conditions more adequately met?

Part of Quiz 2

(e) Which model would you recommend and why?

(f) Which model form makes the most sense in context? Explain.

Notes

- For a formal test of normality of the residuals, you can use something like shapiro.test(resid(model1)), but many experts prefer the visual inspection of the graphs
- For an outlier test, you can try something like library(car); outlierTest(model1), but this is mostly for flagging unusual observations. You still need to investigate whether you have any justification for removing them or treating them differently.
- Residual standard error goes by many names, including root mean square error, $\hat{\sigma}$
- The fact that we cannot interpret the intercept here (was no race run in the year 0) is not necessarily a problem.