Assumptions of Linear Regression

- Existence: for any fixed value of X, Y has a probability distribution with finite mean and variance.
- Independence: Y values are statistically independent
  - violated if repeated measurements are taken on the same individual
- Linearity: the mean value of $Y|X$ is a straight-line function of X
- Homoscedasticity: the conditional distributions of $Y|X$ have the same variance at all levels of X
  - $\sigma^2_{Y|X} = \sigma^2$ for all X
- Normality: the distribution of the errors is normal with mean 0

Checking Linearity

Do the means of the conditional distributions lie on a line?

to check: (A) Scatter diagram of $Y_i$ and $X_i$
(B) Scatter diagram of residuals $\hat{e}_i$ vs. $X_i$ or $\hat{Y}_i$
Homoscedasticity (Equal variance of $Y$'s at each value of $X$)

to check:

1. scatter diagram of $Y_i$ and $X_i$
2. residual plot of $\hat{e}_i$ vs $X_i$ or vs predicted values
3. studentized residuals easier to interpret

if we generate “studentized residuals”

$$e^*_i = \frac{\hat{e}_i}{\text{stderr}(\hat{e}_i)}$$

then the studentized errors $e^*_i$ are approximately normally distributed with mean 0 and variance 1!

---

I created studentized residuals by using the “Vars” menu on the “Fit” window. Then I did a scatterplot of the studentized residuals vs. predicted values.

---

```
proc reg data = oecdold graphics ;
model pch = pcgdp / r ;
id pcgdp ;
plot pch * pcgdp pred. * pcgdp / overlay ;
run ;
plot residual. * pred. ;
run ;
```
Normality (The errors $e_i$ are normally distributed.)

to check: The residuals $\hat{e}_i$ are our estimates of the errors. Use plots or summary statistics to check their normality.

I got this plot by going into the “Graphs” menu on the “Fit” window in Insight.

Can we find a transformation that will improve homoscedasticity and normality without making linearity any worse? Let’s try log 10 first. Why was this a reasonable choice?
The log transformation was too strong. Let’s try square root, which is in between no transformation and log.

Scatter plot and regression line

Plot of residuals vs. predicted values
Normal qq plot of residuals

![Normal qq plot of residuals](image)

This looks as good as we can hope for on all counts.

Now let's transform the predicted values of square root of PCH back to the untransformed units.

I created a new variable, the square of the predicted values from the regression of sqrt(pch) on pcgdp. I then plotted it against pcgdp.

When we have transformed back to the original scale, does the relationship between \( pch \) and \( pcgdp \) appear to be linear or curvilinear?

Do predicted values of \( pch \) change more rapidly for a change of one unit in \( pcgdp \) in lower or higher ranges of \( pcgdp \)?

<table>
<thead>
<tr>
<th>country</th>
<th>pch</th>
<th>origpred</th>
<th>pcgdp</th>
</tr>
</thead>
<tbody>
<tr>
<td>united_states</td>
<td>2051</td>
<td>1683.412</td>
<td>18.1429</td>
</tr>
<tr>
<td>canada</td>
<td>1483</td>
<td>1550.555</td>
<td>17.2857</td>
</tr>
<tr>
<td>iceland</td>
<td>1241</td>
<td>1301.237</td>
<td>15.6714</td>
</tr>
<tr>
<td>sweden</td>
<td>1233</td>
<td>1073.757</td>
<td>13.8571</td>
</tr>
<tr>
<td>switzerland</td>
<td>1225</td>
<td>1341.271</td>
<td>15.8571</td>
</tr>
<tr>
<td>norway</td>
<td>1149</td>
<td>1301.237</td>
<td>15.6714</td>
</tr>
<tr>
<td>france</td>
<td>1105</td>
<td>884.4225</td>
<td>12.2857</td>
</tr>
<tr>
<td>germany</td>
<td>1093</td>
<td>1020.308</td>
<td>13.4286</td>
</tr>
<tr>
<td>luxemborg</td>
<td>1050</td>
<td>1203.798</td>
<td>14.8571</td>
</tr>
<tr>
<td>netherlands</td>
<td>1041</td>
<td>968.212</td>
<td>13</td>
</tr>
<tr>
<td>austria</td>
<td>982</td>
<td>835.9666</td>
<td>11.8571</td>
</tr>
<tr>
<td>finland</td>
<td>949</td>
<td>951.1459</td>
<td>12.8571</td>
</tr>
<tr>
<td>australia</td>
<td>939</td>
<td>884.4225</td>
<td>12.2857</td>
</tr>
<tr>
<td>japan</td>
<td>915</td>
<td>1020.308</td>
<td>13.4286</td>
</tr>
<tr>
<td>belgium</td>
<td>879</td>
<td>835.9666</td>
<td>11.8571</td>
</tr>
<tr>
<td>italy</td>
<td>841</td>
<td>900.8819</td>
<td>12.4286</td>
</tr>
<tr>
<td>denmark</td>
<td>792</td>
<td>1037.969</td>
<td>13.5714</td>
</tr>
<tr>
<td>united_kingdom</td>
<td>758</td>
<td>900.8819</td>
<td>12.4286</td>
</tr>
<tr>
<td>new_zealand</td>
<td>733</td>
<td>728.2181</td>
<td>10.8571</td>
</tr>
<tr>
<td>ireland</td>
<td>561</td>
<td>449.5591</td>
<td>7.8571</td>
</tr>
<tr>
<td>spain</td>
<td>521</td>
<td>535.0143</td>
<td>8.8571</td>
</tr>
<tr>
<td>portugal</td>
<td>386</td>
<td>350.6082</td>
<td>6.5714</td>
</tr>
<tr>
<td>greece</td>
<td>337</td>
<td>340.3758</td>
<td>6.4286</td>
</tr>
<tr>
<td>turkey</td>
<td>148</td>
<td>212.9888</td>
<td>4.4286</td>
</tr>
</tbody>
</table>
Note on Transformations:
All the transformations we are using are “monotonic,” meaning that changes in the untransformed value correspond to changes in the same direction in the transformed value. Thus, for example, if we find that
$$\sqrt{Y} = \beta_0 + \beta_1 X$$
and $\beta_1$ is positive and the relationship is significant,
then we can say also that $Y$ will increase as $X$ increases (just not in a straight line). If $\beta_1$ is negative, we know $Y$ will decrease with increasing $X$.

The previous methods of verifying the assumptions of a linear regression model can be summarized as follows:
1. check the marginal (univariate) distributions of $X$ and $Y$ to get some information about the data
2. produce a scatterplot of $Y$ vs $X$ to investigate the relationship and see if it seems linear
3. (fit the regression model) and generate residuals
4. construct graphical summaries or numerical summaries of the residuals to check the normality assumption. Standardized residuals may be easier to interpret.
5. construct plots of the residuals $e_i$ vs $X_i$ or $\hat{Y}_i$. These can be used to investigate both linearity and homoscedasticity.

What to do when assumptions are violated:

Lack of linearity
- transform data (square root, log, reciprocal)
- add more independent variables
- eliminate outliers

Lack of homoscedasticity
- transform data
  - log and square root stabilize variance when it increases as $Y$ increases

Lack of normality
- transform data
- use a non-parametric method (Kruskal-Wallis)
- eliminate outliers

Choosing the best transformation
- If the original data satisfy the assumptions of linear regression, don’t transform!
- If not, try several transformations to see which one (or ones) comes closest to satisfying the assumptions.
- If more than one transformation appears to satisfy the assumptions, chose the one with the highest $R^2$
  - $t$ test
  - overall $F$ test
What to do about outliers

- Make sure value is not an error
- Determine (if possible) whether there is some reason why this observation would be expected to be different from the others—may suggest future studies
- Fit the regression line with and without the outlier and see if the overall conclusions remain the same “influential points”
- Transform data?
- Consider research goals

Interpreting the Final Model

- What is the interpretation of the estimated slope?
- Is the association positive or negative?
- Does this make sense intuitively, based on what the data represent?
- What other variables could be confounders?
- Are there other analyses that you might consider doing? New questions raised?