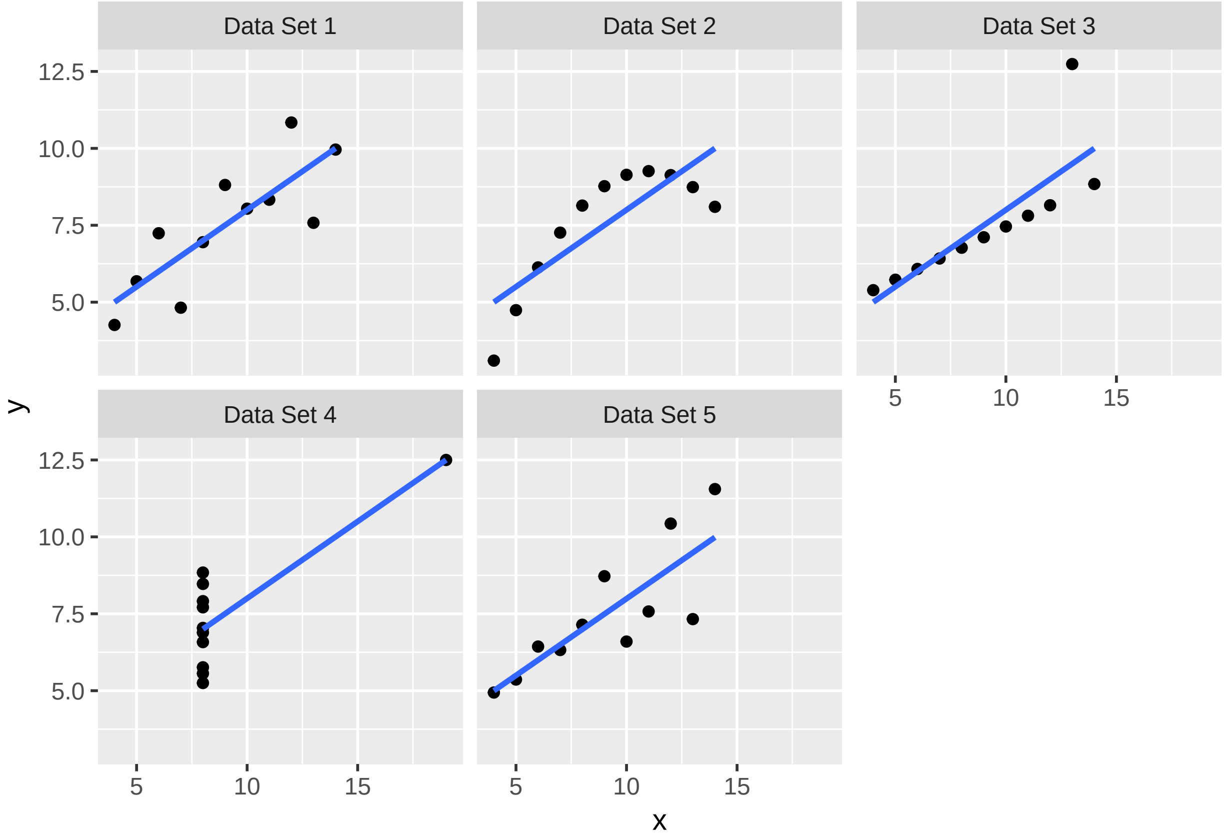
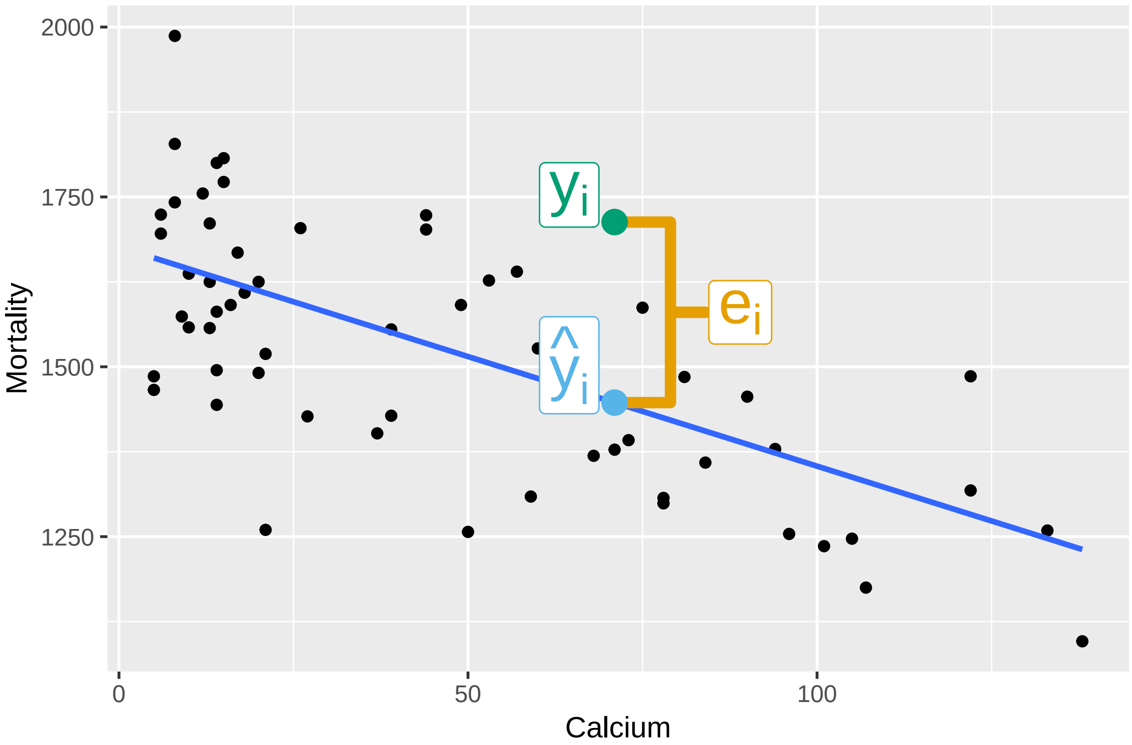
Linear Regression: Conditions for Inference, Residual Diagnostics

All 5 Have Essentially the Same Estimated Intercept, Slope, R^2, and Residual Standard Deviation!



- Briefly, **conditions for linear regression** (see last page for more detail):
  - Sample **representative** of population
  - No **outliers** (points that don't fit the trend)
  - **Linear** relationship
  - **Independent** observations
  - **Normally** distributed residuals (or large enough sample size)
  - **Equal variability** of residuals
- **Use plots** to help diagnose the appropriateness of a linear model:
  - Scatter plot of explanatory (x axis) vs. response (y axis)
  - Scatter plot of predicted (x axis) vs. residual (y axis)
  - Histogram or density plot of residuals (x axis)
- Checks of whether the sample is representative and whether the observations are independent come from thinking about data collection process, not plots.

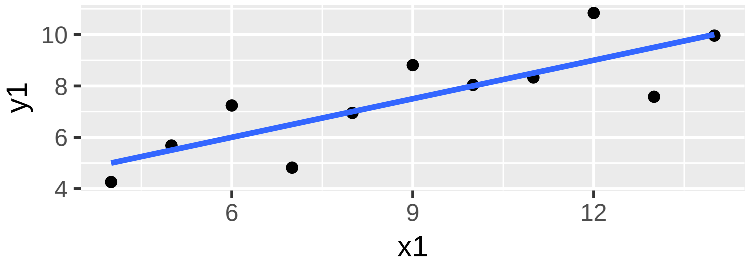
A Reminder about Residuals



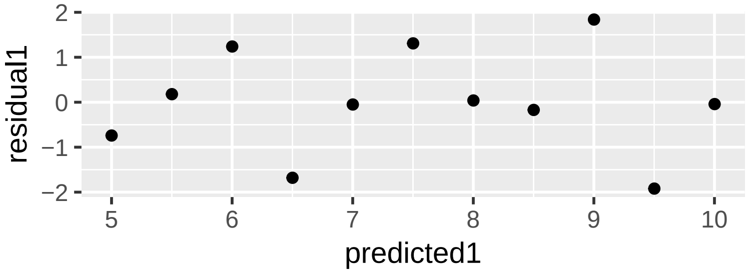
- Residuals give the vertical distance between a data point and the line of best fit
- Positive if point above line, negative otherwise
- Residual = *Observed* - *Predicted*
- $e_i = y_i - \hat{y}_i$ (e stands for error)

Anscombe Quintet: Data Set 1 (All Is Well)

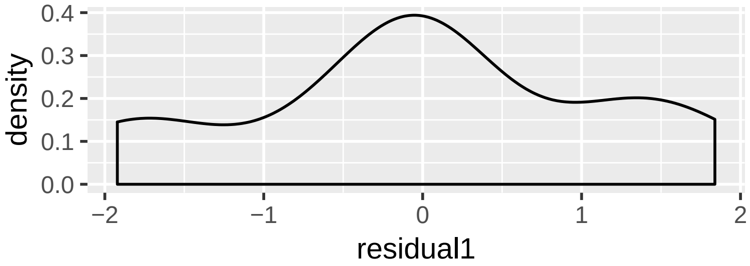
```r
ggplot(data = anscombe, mapping = aes(x = x1, y = y1)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



```r
linear_fit1 <- lm(y1 ~ x1, data = anscombe)
anscombe <- anscombe %>% mutate(
  predicted1 = predict(linear_fit1),
  residual1 = residuals(linear_fit1)
)
ggplot(data = anscombe, mapping = aes(x = predicted1, y = residual1)) +
  geom_point()
```



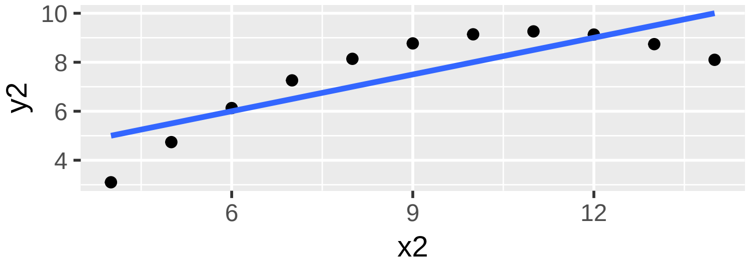
```r
ggplot(data = anscombe, mapping = aes(x = residual1)) +
  geom_density()
```



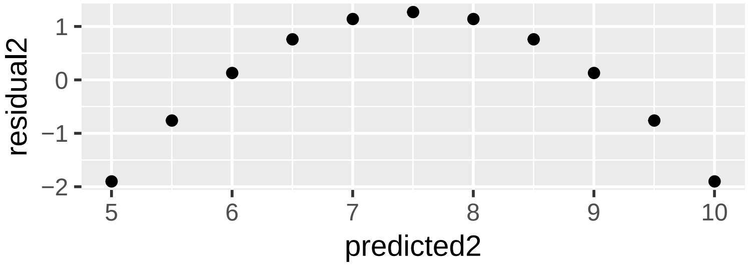
- **Outliers**? No
- **Linear** relationship? Yes
- **Normally** distributed residuals? Good enough
- **Equal variability** of residuals? Yes

Anscombe Quintet: Data Set 2 (Nonlinear)

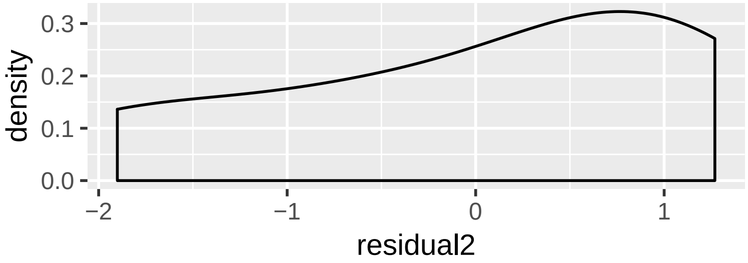
```
ggplot(data = anscombe, mapping = aes(x = x2, y = y2)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



```
linear_fit2 <- lm(y2 ~ x2, data = anscombe)
anscombe <- anscombe %>% mutate(
  predicted2 = predict(linear_fit2),
  residual2 = residuals(linear_fit2)
)
ggplot(data = anscombe, mapping = aes(x = predicted2, y = residual2)) +
  geom_point()
```



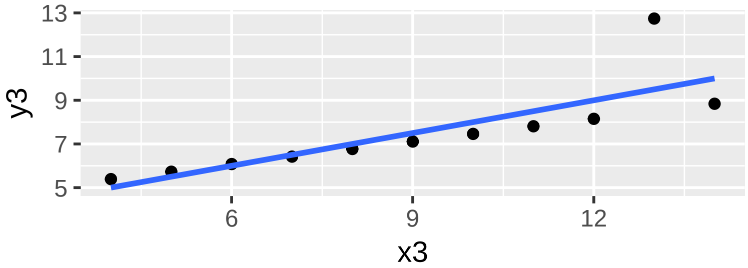
```
ggplot(data = anscombe, mapping = aes(x = residual2)) +
  geom_density()
```



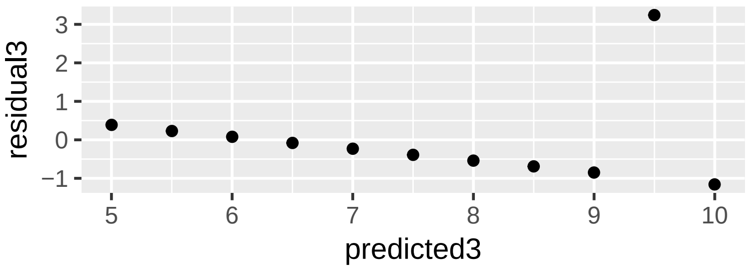
- **Outliers**? No
- **Linear** relationship? No (**this is a problem!!**)
- **Normally** distributed residuals? No perfect, probably good enough
- **Equal variability** of residuals? Yes

Anscombe Quintet: Data Set 3 (Outlier)

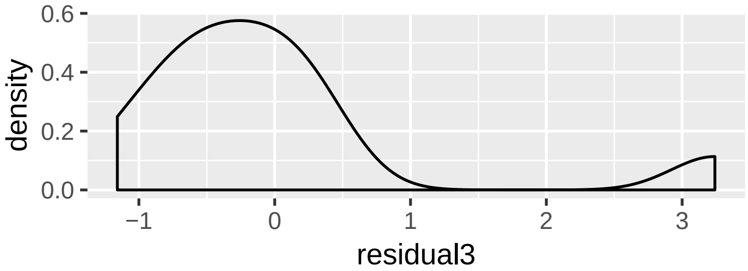
```
ggplot(data = anscombe, mapping = aes(x = x3, y = y3)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



```
linear_fit3 <- lm(y3 ~ x3, data = anscombe)
anscombe <- anscombe %>% mutate(
  predicted3 = predict(linear_fit3),
  residual3 = residuals(linear_fit3)
)
ggplot(data = anscombe, mapping = aes(x = predicted3, y = residual3)) +
  geom_point()
```



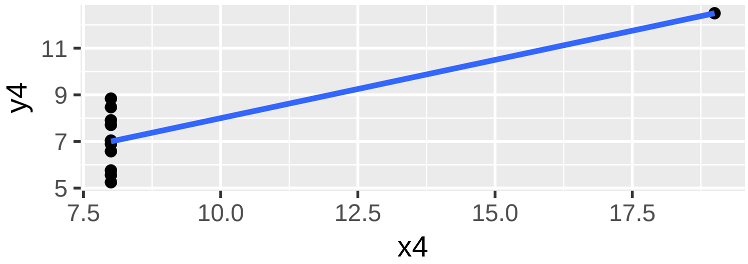
```
ggplot(data = anscombe, mapping = aes(x = residual3)) +
  geom_density()
```



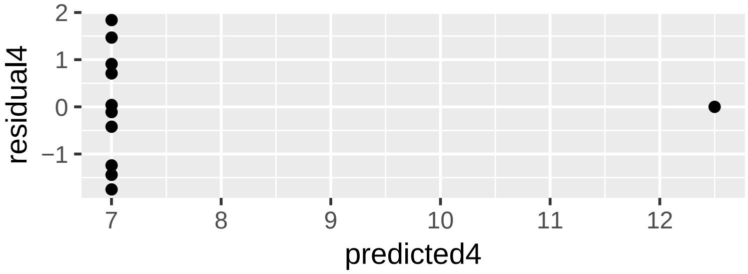
- **Outliers**? Yes (**this is a problem!!**)
- **Linear** relationship? Yes (other than the outlier)
- **Normally** distributed residuals? No, there is an outlier
- **Equal variability** of residuals? Yes (other than the outlier)

Anscombe Quintet: Data Set 4 (Outlier)

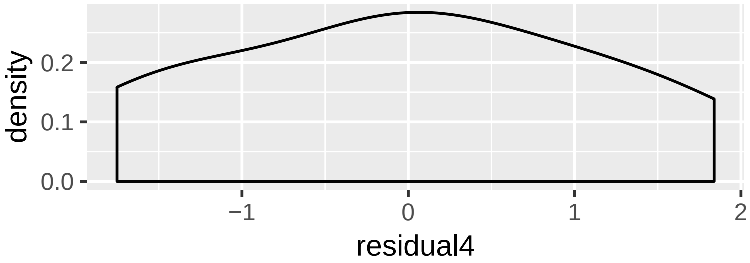
```
ggplot(data = anscombe, mapping = aes(x = x4, y = y4)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



```
linear_fit4 <- lm(y4 ~ x4, data = anscombe)
anscombe <- anscombe %>% mutate(
  predicted4 = predict(linear_fit4),
  residual4 = residuals(linear_fit4)
)
ggplot(data = anscombe, mapping = aes(x = predicted4, y = residual4)) +
  geom_point()
```



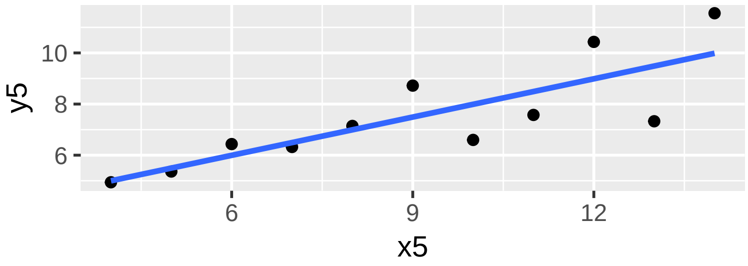
```
ggplot(data = anscombe, mapping = aes(x = residual4)) +
  geom_density()
```



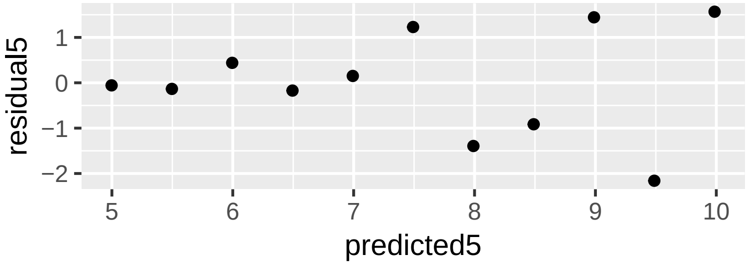
- **Outliers**? Yes (**this is a problem!!**)
- **Linear** relationship? Difficult to assess
- **Normally** distributed residuals? Yes
- **Equal variability** of residuals? Difficult to assess

Anscombe Quintet: Data Set 5 (Lack of Equal Variability of Residuals)

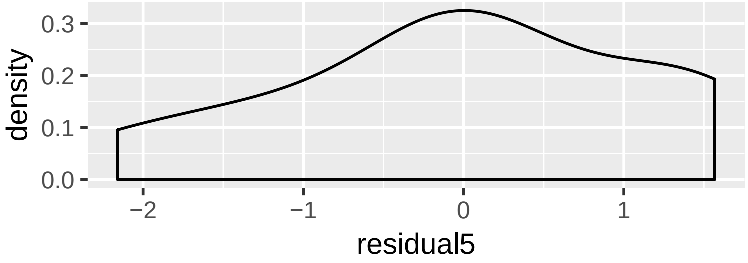
```
ggplot(data = anscombe, mapping = aes(x = x5, y = y5)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



```
linear_fit5 <- lm(y5 ~ x5, data = anscombe)
anscombe <- anscombe %>% mutate(
  predicted5 = predict(linear_fit5),
  residual5 = residuals(linear_fit5)
)
ggplot(data = anscombe, mapping = aes(x = predicted5, y = residual5)) +
  geom_point()
```



```
ggplot(data = anscombe, mapping = aes(x = residual5)) + geom_density()
```



- **Outliers**? No
- **Linear** relationship? Yes
- **Normally** distributed residuals? Yes
- **Equal variability** of residuals? No (**this is a problem!!**)

Regression Conditions

Think of a helpful leprechaun named **R**obert **O'Line**:

- Sample **representative** of population
- No **outliers** (points that don't fit the trend)
- **Linear** relationship
- **Independent** observations
- **Normally** distributed residuals (or large enough sample size)
- **Equal variability** of residuals

Condition	How Important?	How to Check?
Representative	Critical	Think about data collection (randomization?)
No **O**utliers	Very Important	• Scatter Plot of explanatory variable vs response variable • Scatter plot of predicted value vs residuals • histogram or density plot of residuals
Linear relationship	Very Important	• Scatter Plot of explanatory variable vs response variable (pattern is linear) • Scatter plot of predicted value vs residuals (no curved patterns)
Independent observations	Very Important	Think about data collection (randomization?) Situations where observations are **not** independent: • Observations collected over time (e.g., monthly unemployment measurements over time) • Multiple observations on the same person (e.g., baseline and follow-up measurements of health in a clinical trial)
Normally distributed residuals	Somewhat Important	• histogram or density plot of residuals (unimodal, approximately symmetric, no outliers) • ...or large enough sample size
Equal variability of residuals	Somewhat Important	• Scatter Plot of explanatory variable vs response variable (same amount of vertical spread around line for all values of x) • Scatter plot of predicted value vs residuals (same amount of vertical spread for all values of x)