Announcements

- Make sure to sign in on the google form (linked here)
- Midterm October 11

Slope independent of outcome mean

1. Find the distribution of $\bar{Y}$. Recall that $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \beta_0 + \beta_1 X_i + \epsilon_i = \beta_0 + \beta_1 \bar{X} + \bar{\epsilon}$$

The only piece of this that is random is $\bar{\epsilon} \sim \mathcal{N}(0, \sigma^2/n)$, so $\bar{Y} \sim \mathcal{N}(\beta_0 + \beta_1 \bar{X}, \sigma^2/n)$.

2. Show that $E(\bar{Y}\hat{\beta}_1) = E(\bar{Y})E(\hat{\beta}_1)$. Recall that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Using the fact that the X_i are fixed,

$$\begin{aligned}
E(\bar{Y}\hat{\beta}_1) &= E\left((\beta_0 + \beta_1\bar{X} + \bar{\epsilon})\hat{\beta}_1\right) \\
&= \beta_0\beta_1 + \beta_1^2\bar{X} + E\left(\bar{\epsilon}\frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) \\
&= \beta_0\beta_1 + \beta_1^2\bar{X} + E\left(\bar{\epsilon}\frac{\sum_{i=1}^n (X_i - \bar{X})(\beta_0 + \beta_1 X_i + \epsilon_i - \beta_0 - \beta_1\bar{X} - \bar{\epsilon})}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) \\
&= \beta_0\beta_1 + \beta_1^2\bar{X} + E\left(\bar{\epsilon}\frac{\sum_{i=1}^n (X_i - \bar{X})\beta_0}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) + E\left(\bar{\epsilon}\frac{\sum_{i=1}^n (X_i - \bar{X})\beta_1(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) + E\left(\bar{\epsilon}\frac{\sum_{i=1}^n (X_i - \bar{X})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) \\
&= \beta_0\beta_1 + \beta_1^2\bar{X} + 0 + 0 + \frac{\sum_{i=1}^n (X_i - \bar{X})[E(\epsilon_i\bar{\epsilon}) - E(\bar{\epsilon}^2)]}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \beta_0\beta_1 + \beta_1^2\bar{X} + \frac{\sum_{i=1}^n (X_i - \bar{X})[E(\frac{1}{n}\epsilon_i\sum_{j=1}^n \epsilon_j) - E(\bar{\epsilon}^2)]}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \beta_0\beta_1 + \beta_1^2\bar{X} + \frac{\sum_{i=1}^n (X_i - \bar{X})[\frac{1}{n}E(\epsilon_i^2) - E(\bar{\epsilon}^2)]}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \beta_0\beta_1 + \beta_1^2\bar{X} + \frac{\sum_{i=1}^n (X_i - \bar{X})[\sigma^2/n - \sigma^2/n]}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \beta_0\beta_1 + \beta_1^2\bar{X}
\end{aligned}$$

3. Find the covariance of $\bar{Y}$ and $\hat{\beta}_1$.

$\text{Cov} = E(\bar{Y}\hat{\beta}_1) - E(\bar{Y})E(\hat{\beta}_1) = 0$

4. Apply 7.5.7 from the Stat 110 textbook to show that $\bar{Y}$ and $\hat{\beta}_1$ are independent.

$(\bar{Y}, \hat{\beta}_1)$ is bivariate normal, and within a MVN vector, uncorrelated implies independent.

Redundant summary information

Here's a bunch of useful information (also available here, but be careful of what they call p):

Definitions:

- Sum of squares model (SSM): $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
- Sum of squares error (SSE): $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- Sum of squares total (SST): $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- Degrees of freedom for model with p predictors and an intercept (df_M): p
- Degrees of freedom for error with p predictors and an intercept (df_E): $n - p - 1$
- Residual standard error: $\sqrt{\text{SSE}/\text{df}_E}$
- R^2: $1 - \text{SSE}/\text{SST}$
- Adjusted R^2: $1 - (1 - R^2)\frac{n-1}{\text{df}_E}$

Facts:

- $\text{SSE} + \text{SSM} = \text{SST}$
- $\hat{\sigma}^2 = \text{SSE}/\text{df}_E$
- Under the null (all coefficients are 0),

$$\frac{\text{SSM}/\text{df}_M}{\text{SSE}/\text{df}_E} \sim F_{\text{df}_M, \text{df}_E}$$

```
Call:
lm(formula = spi_ospi ~ mad_gdppc, data = countries)

Residuals:
     Min       1Q   Median       3Q      Max
-61.1775  -7.3751   2.7922   9.2159  15.7948

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 5.9866e+01 1.2971e+00
mad_gdppc   4.6565e-04 4.6213e-05
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.745 on 151 degrees of freedom
  (41 observations deleted due to missingness)
Multiple R-squared:  0.40205,   Adjusted R-squared:
F-statistic:        on   and 151 DF,  p-value:
```

Figure 1: Lm output with missing information

From the partial output above, calculate the following:

1. How many non-NA data points were included.

$n = \text{df}_E + p + 1 = \text{df}_E + 2 = 153$

2. The t-statistics for the intercept and `mad_gdppc` coefficient.

$t = \frac{\text{Estimate}}{\text{Standard error}}$, so $t_{\beta_0} = 59.866/1.2971 = 46.154$ and $t_{\beta_1} = 0.00046565/0.00004621 = 10.08$

3. The p-values of the two t-tests for the intercept and `mad_gdppc` coefficient being 0.

```
# Intercept
1 - pt(46.154, df = 151)
```

```
## [1] 0
```

```
# mad_gdppc coefficient
1 - pt(10.08, df = 151)
```

```
## [1] 0
```

4. A 95% confidence interval for the `mad_gdppc` coefficient.

```
qt(0.975, 151)
```

```
## [1] 1.975799
```

$\hat{\beta}_1 \pm t^* \cdot \text{Standard Error} = 0.00046565 \pm 1.976 \cdot 0.00004621 = (0.000374, 0.000557)$ which doesn't include 0 as expected.

5. The adjusted R^2.

$1 - (1 - R^2)\frac{n-1}{\text{df}_E} = 1 - (1 - 0.402)\frac{152}{151} = 0.398$

6. The sum of squares error, the sum of squares total, and the sum of squares model.

$$\text{SSE} = \text{Residual standard error}^2 \cdot \text{df}_E = 11.745^2 \cdot 151 = 20829.70$$
$$\text{SST} = \frac{\text{SSE}}{1 - R^2} = 20829.70/0.59795 = 34835.19$$
$$\text{SSM} = \text{SST} - \text{SSE} = 14005.49$$

7. The f-statistic and p-value for the test that all coefficients are equal to 0.

$$f = \frac{\text{SSM}/\text{df}_M}{\text{SSE}/\text{df}_E} = \frac{14005.49/1}{20829.70/151} = 101.53$$

```
1 - pf(101.53, 1, 151)
```

```
## [1] 0
```

8. Note that the hypothesis tested in 7 ($H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$) was the same as one of the hypotheses tested in 2. If our framework is consistent, these should give the same answer. Recall from week 2's section that if $T_n \sim t_n$, $T_n^2 \sim F_{1,n}$. Show (numerically) that your calculated t statistic squared is your f statistic, and explain how this shows that the two tests are the same. (Note that this only works because we have a single predictor.)

The two test statistics are within rounding error of each other: $t^2 = 10.08^2 = 101.61 \approx 101.53 = f$. Under the null, the t-statistic T_n of β_1 has a t_n distribution, so T_n^2 will have a $F_{1,n}$ distribution, so with the observed t-statistic t_n and $f = t_n^2$, $P(|t_n| \geq |T_n|) = P(t_n^2 \geq T_n^2) = P(t_n^2 \geq F_{1,n}) = P(f \geq F_{1,n})$.

The linear model for the image is from this chunk of code.

```
countries <- read.csv("data/countries.csv")

# Show n
sum(!is.na(countries$mad_gdppc) & !is.na(countries$spi_ospi))
```

```
## [1] 153
```

```
# Display model
summary(lm(spi_ospi ~ mad_gdppc, countries))
```

```
##
## Call:
## lm(formula = spi_ospi ~ mad_gdppc, data = countries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.177  -7.375   2.792   9.216  15.795
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.987e+01  1.297e+00   46.15   <2e-16 ***
## mad_gdppc   4.657e-04  4.621e-05   10.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.74 on 151 degrees of freedom
##   (41 observations deleted due to missingness)
## Multiple R-squared:  0.402,  Adjusted R-squared:  0.3981
## F-statistic: 101.5 on 1 and 151 DF,  p-value: < 2.2e-16
```

Regression on real data

This section will deal with a data set of country-level statistics from this source with an explanation of the data encoding found here.

```
countries <- read.csv("data/countries.csv")
```

1. Fit a linear model to predict the percent of individuals using the internet in a country (`wdi_internet`) from the log of its GDP per capita (`mad_gdppc`), and formally test whether this association is significant. Provide a visual to support your conclusion.

We want to test $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$ where β_1 is the association between log GDP per capita and percent of individuals with access to internet in a country. We get a slope of 20.97 and a t-statistic of 24.7 for that slope with a p-value of less than $2.2 \times 10^{-16} < \alpha = 0.05$, so we reject the null and conclude there is an association between log GDP per capita and percent of individuals with access to internet in a country (go figure!).

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

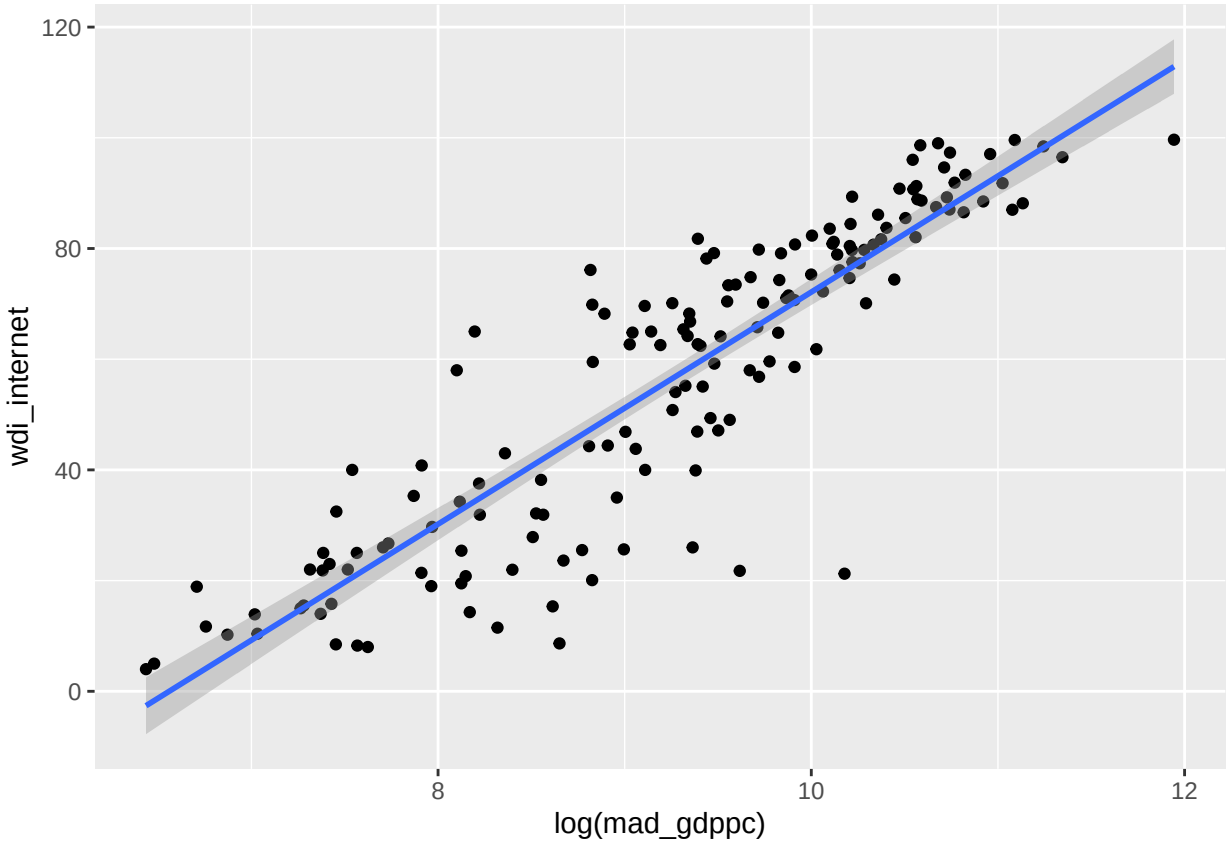
```
lm1 <- lm(wdi_internet ~ log(mad_gdppc), countries)
summary(lm1)
```

```
##
## Call:
## lm(formula = wdi_internet ~ log(mad_gdppc), data = countries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.609  -7.031   1.913   7.592  30.682
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -137.5251     7.9227  -17.36   <2e-16 ***
## log(mad_gdppc)   20.9652     0.8487   24.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.66 on 156 degrees of freedom
##   (36 observations deleted due to missingness)
## Multiple R-squared:  0.7964, Adjusted R-squared:  0.7951
## F-statistic: 610.3 on 1 and 156 DF,  p-value: < 2.2e-16
```

```
ggplot(countries, aes(x=log(mad_gdppc), y=wdi_internet)) +
  geom_point() +
  geom_smooth(method='lm', formula= y~x)
```

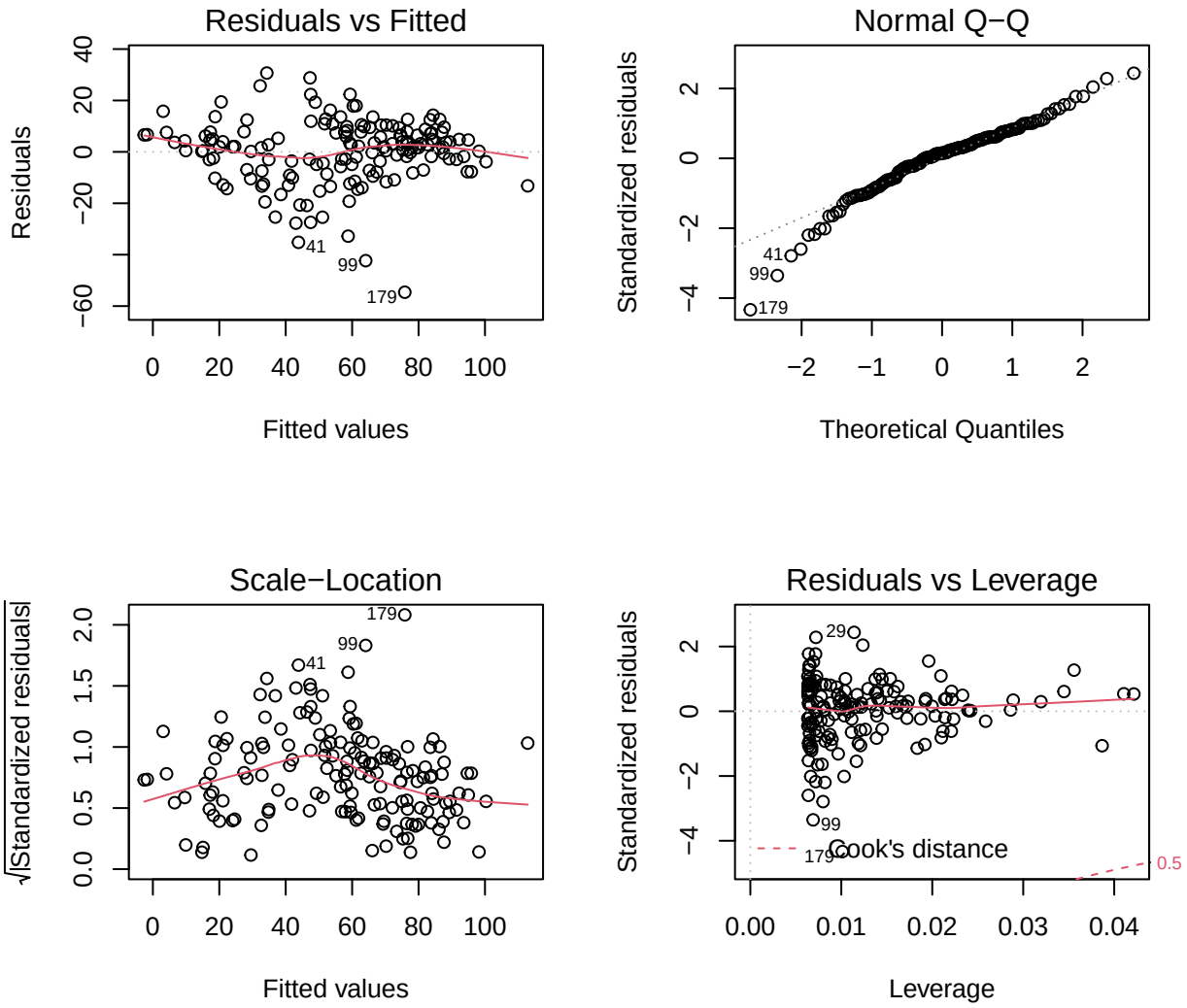
```
## Warning: Removed 36 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 36 rows containing missing values (geom_point).
```



2. Check the assumptions of the model.

```
par(mfrow=c(2,2))
plot(lm1)
```



- Linearity: The Residuals vs Fitted plot shows that there is no clear pattern to the residuals, so linearity is likely upheld.
- Constant variance: Based on the Scale-Location plot, there might be slightly more variance in residuals for countries with GDPs near the world-wide median, but the variance is about constant. (The Residuals vs Fitted plot makes it look like there is more variance in the middle, but note that there's also more data there in the first place.)
- Normality: The Q-Q plot show that the lower tails are slightly lower than expected with the normal assumption, but overall the normal assumption fits very well.
- Independence: This is questionable: even given GDP, it's possible that internet use in a region is correlated because companies able to set up and maintain the infrastructure might work across multiple countries in a region.

3. Uganda has a GDP per capita listed but no statistic for internet access. Provide a point estimate and 90% prediction interval.

```
log(countries[countries$cname=="Uganda",]$mad_gdppc)
```

```
## [1] 7.623359
```

```
predict(lm1, newdata=countries[countries$cname=="Uganda",],
        interval = c("prediction"), level = 0.90)
```

```
##          fit      lwr      upr
## 181 22.30046 1.155289 43.44564
```

How bad are correlated residuals?

Let $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ with marginal $\epsilon_i \sim \mathcal{N}(0,1)$ and $\text{Corr}(\epsilon_i, \epsilon_{i+1}) = \rho$ for $i \in \{1, ..., n-1\}$ and $\text{Corr}(\epsilon_i, \epsilon_{i-1}) = \rho$ for $i \in \{2, ..., n\}$ and $\text{Corr}(\epsilon_i, \epsilon_j) = 0$ otherwise. Write a function to use simulation to find the probability of rejecting the null $H_0 : \beta_1 = 0$, the expected value $E(\hat{\beta}_1)$, and the standard deviation $\text{SD}(\hat{\beta}_1)$ in the following situations:

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.1.3
```

```
nsims = 1000
n = 10
b0 = 1

run_sim = function(nsims, n, p, b0, b1, sorted=FALSE) {
  # Covariance matrix
  Sigma = matrix(0, nrow = n, ncol = n)
  diag(Sigma) <- 1
  for (i in 2:n) {
    Sigma[i, i-1] <- p
    Sigma[i-1, i] <- p
  }

  pval = vector(length = nsims)
  coef = vector(length = nsims)

  for (i in 1:nsims) {
    # Generate x
    if (sorted) {
      x <- sort(rgamma(n, 3, 2/5))
    } else {
      x <- rgamma(n, 3, 2/5)
    }
    # Generate y with multivariate normal
    y <- b0 + b1 * x + mvrnorm(n = 1, rep(0, n), Sigma)

    # Make linear model
    tmp <- summary(lm(y~x))
    pval[i] <- tmp$coefficients[2,'Pr(>|t|)']
```

```
    coef[i] <- tmp$coefficients[2,'Estimate']
  }

  return(c("signif" = mean(pval < 0.05), "mean" = mean(coef), "sd" = sd(coef)))
}
```

1. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$, $\rho = 0$, $\beta_0 = 1$, $\beta_1 = 1$.

```
run_sim(nsims, n, 0, b0, 1, sorted=FALSE)
```

```
##     signif       mean         sd
## 1.00000000 0.99724913 0.08815965
```

2. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$, $\rho = 0.5$, $\beta_0 = 1$, $\beta_1 = 1$.

```
run_sim(nsims, n, 0.5, b0, 1, sorted=FALSE)
```

```
##     signif       mean         sd
## 1.00000000 1.00415415 0.09289591
```

3. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$ sorted, $\rho = 0$, $\beta_0 = 1$, $\beta_1 = 1$.

```
run_sim(nsims, n, 0, b0, 1, sorted=TRUE)
```

```
##     signif       mean         sd
## 1.00000000 0.99890711 0.09075181
```

4. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$ sorted, $\rho = 0.5$, $\beta_0 = 1$, $\beta_1 = 1$.

```
run_sim(nsims, n, 0.5, b0, 1, sorted=TRUE)
```

```
##    signif      mean        sd
## 0.9990000 0.9942999 0.1148095
```

5. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$, $\rho = 0$, $\beta_0 = 1$, $\beta_1 = 0$.

```
run_sim(nsims, n, 0, b0, 0, sorted=FALSE)
```

```
##      signif        mean          sd
## 0.053000000 0.001240469 0.095298579
```

6. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$, $\rho = 0.5$, $\beta_0 = 1$, $\beta_1 = 0$.

```
run_sim(nsims, n, 0.5, b0, 0, sorted=FALSE)
```

```
##        signif          mean            sd
##  0.0490000000 -0.0005393505  0.0892190425
```

7. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$ sorted, $\rho = 0$, $\beta_0 = 1$, $\beta_1 = 0$.

```
run_sim(nsims, n, 0, b0, 0, sorted=TRUE)
```

```
##        signif          mean            sd
##  0.0520000000 -0.0009530659  0.0911075034
```

8. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$ sorted, $\rho = 0.5$, $\beta_0 = 1$, $\beta_1 = 0$.

```
run_sim(nsims, n, 0.5, b0, 0, sorted=TRUE)
```

```
##      signif        mean          sd
## 0.135000000 0.002331801 0.116487046
```

9. What conclusions can you draw?

It is very easy for the linear model to pick up significance and estimate the coefficient about right when it is non-zero. Even when the ϵ_i are correlated, if that correlation is not associated with the X_i values, the coefficient is unaffected. The only time the coefficient estimate is affected is when X_i of similar magnitude have ϵ_i that are correlated; this causes groupings of Y_i based on groupings of X_i even when there is no association through β_1. Thus, linear models are quite robust to violations of independence if (and this is often a big "if") the correlation in outcomes is not related to the correlation in predictors.